Desarrollo de una metodología para mapear de un espacio cualitativo a uno cuantitativo empleando la distancias Chi-Square como soporte para tareas de agrupamiento y clasificación.

Tesis presentada como requisito para aspirar al título de Magister en Ingeniería Eléctrica.

Estudiante: Luis Ariosto Serna Cardona

Director: Álvaro Ángel Orozco Gutiérrez.

Universidad Tecnológica de Pereira
Facultadad de Ingeniería - Programa de Ingeniería Eléctrica
Maestría en Ingeniería Eléctrica
Grupo de investigación en Automática
Pereira, Risaralda, Colombia
2021

1. Resumen

El uso recurrente de bases de datos con variables categóricas en diferentes aplicaciones, demanda nuevas alternativas para identificar patrones relevantes. La clasificación y las técnicas de agrupamiento se han convertido en enfoques interesantes para procesar este tipo de datos. Sin embargo, varios métodos propuestos en la literatura tienen inconvenientes para reconocer variables categóricas debido a la complejidad de los datos: estos son, alto costo computacional o bajo rendimiento. Por esta razón, proponemos un enfoque supervisado y otro no supervisado basados en una representación de datos categóricos en un espacio euclidiano utilizando una medida de disimilitud de Chi-cuadrado ($\mathbf{C-S}$). Esto permite entender los datos categóricos como numéricos, haciendo posible el agrupamiento de los datos empleando el algoritmo K-means con la típica distancia euclidiana. Además, aumenta el nivel de acierto de los clasificadores convencionales (clasificador bayesiano lineal **(LDC)**- clasificador bayesiano cuadrático **(QDC)**- K-Nearest Neighbor **(KNN)**- Máquinas de vectores de soporte **(SVM)**) y disminuye el tiempo computacional de los algoritmos al combinarlo con **t-SNE**. Las dos etapas de este proyecto se dividen así: Primero, Comparar el método de agrupamiento propuesto con las técnicas de vanguardia para el aprendizaje no supervisado: agrupamiento basado en estructura **(SBC)**, K modas, disimilitud ponderada, algoritmos Mkm- (nof y ndm). En los cuales, evaluamos la información mutua normalizada **(NMI)**, el índice de ajuste rand **(ARI)** y la precisión **(AC)**. Los resultados muestran, que nuestra propuesta supera a los métodos de comparación en nueve conjuntos de datos categóricos diferentes. Además, el costo computacional es menor para este enfoque que para los otros algoritmos. La segunda etapa, fue enfocada directamente en la clasificación de los datos utilizando 4 clasificadores convencionales. Luego, se empleó la incrustación de vecinos estocásticos que siguen una distribución t- student **(t-SNE)** para reducir la dimensionalidad de los datos a dos o tres características, lo que permite una reducción significativa de los tiempos de cálculo en los métodos de aprendizaje supervisado. Se evaluó el rendimiento del enfoque propuesto en términos de precisión **(AC)** para varias configuraciones experimentales y conjuntos de datos categóricos tomados de UCI Repository. Los resultados muestran que el mapeo con la medida de disimilitud ($\mathbf{C-S}$) y el algoritmo t-SNE disminuyen considerablemente el tiempo computacional en las tareas de reconocimiento, mientras se conserva la precisión. Además, cuando se aplicó el mapeo con ($\mathbf{C-S}$) a los conjuntos de datos, se mejora la separabilidad de las clases, por lo que el rendimiento de los algoritmos aumenta.

Palabras claves: Agrupamiento, Clasificación, K-means, Mapeo, Métrica, t-SNE, Variable cualitativa, Disimilitud, Chi-cuadrada.

2. Abstract

The recurrent use of databases with categorical variables in different applications demands new alternatives to identify relevant patterns. Classification and clustering techniques have become interesting approaches to processing this type of data. However, several methods proposed in the literature have disadvantages to recognize categorical variables because of the complexity of the data: these are high computational cost or low performance. For this reason, we propose a supervised and an unsupervised approach based on a representation of categorical data in a Euclidean space using a Chi-square measure of dissimilarity. This allows us to understand categorical and numerical data, making it possible to group the data using the K-means algorithm with the typical Euclidean distance. In addition, it increases the level of success of conventional classifiers **(LDC-QDC-KNN-SVM)** and reduces the computational time of the algorithms when combined with **t-SNE**. The two stages of this project are divide: First, Compare the proposed clustering method with state-of-the-art techniques for unsupervised learning, such as: structure-based clustering **(SBC)**, K modes, W-dissimilarity, Mkm-nof and Mkm-ndm algorithms. In which we test the normalized mutual information **(NMI)**, adjusted rand index **(ARI)** and the accuracy **(AC)**. results that are surpassed with our proposal in the nine databases. Also, the computational time is lower for our approach than the other algorithms. The second stage was directly focused on the classification of the data using 4 conventional classifiers. Then, we employ the **(t-SNE)** for reducing dimensionality of data to two or three features, allowing a significant reduction of computational times in learning methods. We evaluate the performance of proposed approach in terms of accuracy for several experimental configurations and public categorical datasets downloaded for the UCI repository. Results show that $(\mathbf{C}-\mathbf{S})$ mapping and **t-SNE** considerably diminish computational times in recognitions tasks, while the accuracy preserved. Also, when we apply only the $(\mathbf{C}-\mathbf{S})$ mapping to the datasets, the separability of classes is enhanced, thus, the performance of learning algorithms is clearly increased.

Keywords: Cluster, classification, K-means, Mapping, Metric, Qualitative variable, t-SNE, Dissimilarity, Chi-square.

Índice

Índice de figuras

Índice de cuadros

3. Métodos y abreviaciones

En esta sección encontraremos todos los datos, símbolos y abreviaciones que se utilizarán para el desarrollo de esta tesis. En las siguientes tablas encontraremos las técnicas de aprendizaje de máquina utilizadas, los métodos, etc.

Cuadro 1: Lista de kernels usadados en el clasificador SVM.

Kernels	Ecuación
RBF or Gaussian	$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2}\right)$
Linear	$k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^\top \mathbf{x}_2$
polynomial	$k(\mathbf{x}_1, \mathbf{x}_2) = \left(\mathbf{x}_1^\top \mathbf{x}_2 + 1\right)^p$
Sigmoidal	$k(\mathbf{x}_1, \mathbf{x}_2) = \tanh\left(\beta_0 \mathbf{x}_1^\top \mathbf{x}_2 + \beta_1\right)$

Cuadro 2: Abreviación e información relevante de las bases de datos categóricas usadas en la etapa de clasificación descargadas del repositorio publico UCI Repository. [1]

Base de datos	Muestras	Características	clases	Distribución
Audiology (Standardized) (A)	226	69	2	$\{124, 76\}$
Balloons (B)	16	4	2	$\{12, 8\}$
Breast Cancer (diagnosis) (BC)	699	9	2	$\{458, 241\}$
Chess (King-Rook vs. King-Pawn) (C)	3196	36	2	$\{1669, 1527\}$
Lymphography Domain (LD)	148	18	2	$\{81, 61\}$
Molecular Biology (Promoter Gene Sequences) (MB)	106	57	2	$\{53, 53\}$
Congressional Voting Records (V)	435	16	2	$\{267, 168\}$

Cuadro 3: Abreviación e información relevante de las bases de datos categóricas usadas en la etapa de agrupamiento descargadas del repositorio publico UCI Repository. [1]

Base de datos	Muestras	Características	clases	Distribución
Fitting contact lenses (FTL)	24	4	3	$\{4, 5, 15\}$
Ballon (B)	20	4	2	$\{8, 12\}$
Space Shuttle Autolanding (SSA)	15	6	2	$\{6, 9\}$
Soybean-small (SS)	47	35	4	$\{10, 10, 10, 17\}$
Hayes-Roth-Hayes-Roth (HRHR)	132	4	3	$\{51, 51, 30\}$
Lymphography Domain (LD)	142	18	2	$\{81, 61\}$
Vote (V)	435	16	2	$\{168, 267\}$
Breast Cancer (BC)	699	9	2	$\{458, 241\}$
Promoters (P)	106	57	2	$\{53, 53\}$

Cuadro 4: Descripción de los métodos y bases de datos categóricas utilizadas en los algoritmos de clasificación.

Abreviación del experimento	Descripción
(A)c	Database (A) + classifiers
(B)c	Database (B) + classifiers
(BC)c	Database (BC) + classifiers
(C)c	Database (C) + classifiers
(LD)c	Database (LD) + classifiers
(MB)c	Database (MB) + classifiers
(V)c	Database (V) + classifiers
(A) + (C-S)	Database (A) + Chi-Square Mapping + classifiers
(B) + (C-S)	Database (B) Chi-Square Mapping + classifiers
(BC) + (C-S)	Database (BC) + Chi-Square Mapping + classifiers
(C) + (C-S)	Database (C) + Chi-Square Mapping + classifiers
(LD) + (C-S)	Database (LD) + Chi-Square Mapping + classifiers
(MB) + (C-S)	Database (MB) + Chi-Square Mapping + classifiers
(V) + (C-S)	Database (V) + Chi-Square Mapping + classifiers
(A) + (C-S) + (t-SNE)	Database (A) + Chi-Square Mapping + t-SNE + classifiers
(B) + (C-S) + (t-SNE)	Database (B) + Chi-Square Mapping + t-SNE + classifiers
(BC) + (C-S) + (t-SNE)	Database (BC) + Chi-Square Mapping + t-SNE + classifiers
(C) + (C-S) + (t-SNE)	Database (C) + Chi-Square Mapping + t-SNE + classifiers
(LD) + (C-S) + (t-SNE)	Database (LD) + Chi-Square Mapping + t-SNE + classifiers
(MB) + (C-S) + (t-SNE)	Database (MB) + Chi-Square Mapping + t-SNE + classifiers
(V) + (C-S) + (t-SNE)	Database (V) + Chi-Square Mapping + t-SNE + classifiers
(A) + (t-SNE)	Database (A) + t-SNE + classifiers
(B) + (t-SNE)	Database (B) + t-SNE + classifiers
(BC) + (t-SNE)	Database (BC) + t-SNE + classifiers
(C) + (t-SNE)	Database (C) + t-SNE + classifiers
(LD) + (t-SNE)	Database (LD) + t-SNE + classifiers
(MB) + (t-SNE)	Database (MB) + t-SNE + classifiers
(V) + (t-SNE)	Database (V) + t-SNE + classifiers

4. Introducción

Existen dos tipos de datos, cuantitativos y cualitativos. Los datos cuantitativos son continuos en el tiempo, permitiendo un mejor manejo en los algoritmos de aprendizaje supervisado y no supervisado, esto se debe, a que las métricas existentes son las idóneas para datos continuos. Sin embargo, los datos categóricos no tienen esa facilidad, pues son discontinuos en el tiempo, siendo esto uno de los principales problemas a la hora de preprocesar este tipo de datos. Es así, como Mario Tascón [2], detalla la importancia de las grandes cantidades de datos cualitativos en la industria, academia, sector financiero, etc. También, define lo importante que es para todos los sectores, saber utilizar estos datos y sacar el mayor beneficio ya sea económico o de cualquiera otra índole, que se ha facilitado gracias a la industria 4.0. Por ende, el rápido aumento e integración de bases de datos provee a los investigadores e ingenieros de nuevos recursos que pueden ser analizados para hacer descubrimientos científicos, optimizar procesos industriales y encontrar relaciones o patrones entre conjuntos de datos. Los investigadores han establecido algoritmos y han adoptado nuevos métodos para el procesamiento de grandes cantidades de datos (Minería de datos) permitiendo resumir la información en un conjunto mucho más pequeño, preservando la estructura de los datos y destacando las características más relevantes de los mismos [3].

En este orden de ideas, uno de los métodos que hace posible la descripción de los datos es el análisis cluster o clasificación no supervisada, el cual divide los datos en grupos, encontrando una relación inherente entre los datos y sus características, capturando la estructura natural de la información [4]. Un algoritmo de agrupamiento asigna una etiqueta a cada dato, concediéndole así la membresía de pertenencia a un grupo, basándose en la similitud con los demás integrantes del grupo. Uno de los métodos de agrupamiento más comunes y eficientes es el algoritmo K-Means [5–7]; Sin embargo, la elección de una medida de similitud (métrica) generalmente se hace a conveniencia y depende de la aplicación. Esto, debido al tipo de variables que contiene la base de datos influenciando en la elección, ya que no es apropiado calcular la media aritmética de un conjunto de datos con variables de tipo nominal, categóricas o cualitativas, como si fuesen variables cuantitativas [3].

El uso creciente de bases de datos con variables de tipo cualitativo demanda nuevos enfoques a la hora de hacer análisis de grupos debido a los problemas que representan este tipo de datos, como el alto costo computacional, el almacenamiento de memoria del algoritmo y la poca eficiencia a un conjunto masivo de datos. Existen diversos trabajos de agrupamiento que discuten la problemática sobre datos de tipo categórico [8–11], pero ninguno de estos libros ofrece una solución a este problema. Y la recomendación usual de todos estos es binarizar los datos y usar medidas de similitud binarias [12, 13].

El aprendizaje supervisado presenta resultados de clasificación más bajos que los de los algoritmos de aprendizaje no supervisado con este tipo de datos, pues los datos de tipo

categóricos tienen una particularidad y es que están muy traslapados entre sí, y la gran mayoría de veces, se produce una mala clasificación o regresión. Por tal motivo, para algoritmos como SVM [14], GP [15], redes neuronales, etc., se hace difícil un alto acierto de clasificación o regresión, siendo uno de los motivos del poco interés de los investigadores sobre estos temas. De esta manera, se hace pertinente el desarrollo de este proyecto, puesto que el desarrollo de una nueva medida de disimilitud que permita mapear datos de tipo cualitativo a datos cuantitativos, podrá dotar a los investigadores de un nuevo instrumento que les permitirá utilizar las infinidades de algoritmos que existen para datos de tipos cuantitativos. logrando así, un mayor acierto en algoritmos de agrupamiento, clasificación y regresión.

La metodología implementada en este proyecto fue validada con métricas de rendimiento, y bases de datos reales y sintéticas que están totalmente rotuladas y libres de errores. El resto del texto esta organizado de la siguiente manera: el apartado 4.1 Muestra el problema de investigación, los trabajos relacionados y la justificación, el apartado 5 presenta el objetivo general y los objetivos específicos de este trabajo, el apartado 6 Muestra el marco referencial del proyecto, técnicas, teoría y definiciones, el apartado 7, apartado 8, y apartado 9 describe los métodos, el experimento, los resultados y las discusiones de cada objetivo específico, el apartado 10 Presenta las conclusiones, el apartado 11 describe el trabajo a futuro, el apartado 12 expone los resultados académicos de esta investigación, y el apartado 13 muestra los agradecimientos.

4.1. Problema de investigación

La alta demanda en el manejo de datos cuantitativos y cualitativos, hace que empresas, instituciones y todo ente que necesite procesar sus datos, busquen urgentemente un profesional en el tema. Existen varias formas de tratar estas problemáticas, en general se llamará análisis de datos [16]. Hacer un correcto tratamiento de datos requiere tener un conocimiento básico en el tipo de bases de datos que existen, las cuales son nominales y cuantitativas. Al día de hoy, los algoritmos y metodologías más usados para del análisis de datos se enfocan en los datos cuantitativos, ya sea para realizar agrupamiento, regresión, clasificación, etc. En el estado del arte se puede ver claramente la alta demanda para este tipo de bases de datos, contrario es con los datos de tipo categórico debido a que se ha trabajado muy poco con algoritmos sofisticados, como por ejemplo Spectral Clustering [17], SMV [14], GP [15], etc. En la actualidad, una de las técnicas de clasificación más usadas para este tipo de datos son los árboles de decisión, sin embargo, este método carece de robustez y se enfoca principalmente en clasificar la totalidad de los datos, lo que tiene como consecuencia la ineficiencia del método al querer ingresar nuevos datos al modelo, obligando a un reentrenamiento del método. Por otro lado, uno de los métodos de agrupamiento más comunes y eficientes es el algoritmo K-Means [5–7]. Sin embargo, la elección de una métrica generalmente se hace a conveniencia y depende explícitamente de la aplicación que se le dará

al modelo, o sea que depende del tipo de variables que contiene la base de datos. Esto, porque es inapropiado calcular la medida aritmética de un conjunto de datos con variables de tipo nominal., categóricas o cualitativas, como si fuesen variables cuantitativas [3].

Para el etiquetado de datos categóricos se presentan métodos como el Fuzzy C-means (un mejoramiento del algoritmo K-means) [18], pero, tiene muchas dificultades en el tiempo computacional. También, Ralambondrainy [12] presenta un enfoque usando el algoritmo K-Means para agrupar datos de tipo categóricos. Este convierte múltiples atributos categóricos en atributos binarios (1 para la presencia y 0 para la ausencia de dicha categoría), luego trata estos atributos binarios como numéricos en el algoritmo K-Means. Sin embargo, este algoritmo convierte todas las características discretas a tributos binarios, aumentado la dimensionalidad de la basa de datos en (2 elevado al número de características), aumentado el costo computacional y el almacenamiento en memoria del algoritmo. Otros algoritmos como el coeficiente de similitud de Gower [13], medidas de disimilitud, basada en la posición, el alcance y el contenido [19], el algoritmo PAM [8], Clúster jerárquico [20], los algoritmos estadísticos difusos [21] y los métodos de agrupamiento conceptual [22]. Sin embargo, todos estos métodos tienen un problema de eficiencia común cuando son aplicados a una gran cantidad de datos categóricos.

En este mismo orden de ideas, investigaciones recientes para mejorar el rendimiento de los algoritmos de agrupamiento como la de Quian y otros [23], quienes han implementado un algoritmo de agrupamiento el cual se basa en mapear un conjunto de datos categóricos en un espacio euclidiano, su método se basa en la estructura de los datos (SBC) y en experimentos obtuvieron un nivel de acierto de agrupamiento relativamente bueno logrando superar los metodos K-modes [24], Chan [25], Mkm-nof [26] y Mkm-ndm [26]. Sin embargo, este método cuenta con dos problemas, el primero es su elevado costo computacional y el segundo y más relevante, es su pobre nivel de acierto a la hora de agrupar bases de datos categóricas de alta dimensionalidad.

Existen diversos libros de análisis de grupos [8–11] que discuten la problemática de realizar Agrupamiento sobre datos de tipo categórico, pero ninguno de estos libros ofrece una solución al problema. Por ende, la recomendación usual es binarizar los datos y usar medidas de similitud binarias. En otro trabajo Wilson y Martinez [27] realizaron un estudio sobre distancias para datos heterogéneos (bases de datos con variables de tipo cuantitativo y cualitativo mezclados), basados en un enfoque de aprendizaje supervisado donde cada muestra tiene información adicional sobre la clase a la que pertenece. Sin embargo, la precisión del algoritmo es deficiente cuando la base de datos tiene una alta dimensionalidad.

El problema fundamental de esta investigación es la poca capacidad que tienen los algoritmos de aprendizaje supervisado y no supervisado a la hora de agrupar, clasificar y realizar regresión sobre datos de tipo categórico. Es decir. El traslapamiento de los datos en bases categóricas le hace difícil a los algoritmos arrojar resultados con un alto acierto, pues no

existe una métrica adecuada para este tipo de dato. Por ende, el propósito de este proyecto es la implementación de una nueva medida de disimilitud que permita el mapeo de un espacio discreto a uno continúo utilizando una nueva distancia para datos categóricos.

4.2. Trabajos relacionados

En la actualidad es muy común escuchar las palabras Big Data, Ciencia de los datos y Machine Learning. Pero, es aún más común ver a las pequeñas, medianas y grandes empresas producir grandes cantidades de datos y no sacar provecho de ello. Hoy en día podemos ver como Google, Éxito, Amazon, etc., son los reyes del comercio [28], pero todo esto, es gracias a como estas empresas utilizan sus datos, pues son utilizados para predecir los gustos de cada persona, lo que necesita, etc., utilizando Machine Learning. Por ende, es de vital importancia para empresa, entes gubernamentales, instituciones de educación, etc., buscar una manera óptima de procesar datos categóricos.

En este orden de ideas, casi todos los datos que se adquieren son de tipo cualitativo [29], pues se generan en formas de encuestas en hospitales, colegios, universidades y todo tipo de empresas. Por este motivo, se han implementado diferentes ejemplos en la literatura aplicados al procesamiento de datos categóricos. Como el trabajo de la Universidad Nacional de Córdoba [30], en el cual se utilizaron máquinas de soporte vectorial para clasificar de manera confiable a los usuarios de la red eléctrica en consumidores honestos y deshonestos utilizando encuestas con variables sociodemográficas. No obstante, para garantizar una correcta clasificación el usuario debe tener cierto tiempo de afiliación a la compañía. También, investigadores de la facultad de ingeniería eléctrica de la universidad Liubliana, utilizaron redes neuronales para clasificar los perfiles de los consumidores de carga típicos (TLP). Esta investigación [31], mostró la eficiencia que tenía cada cliente y cuánto presupuesto necesita para adquirir la energía necesaria. Sin embargo, este método era bastante complicado de implementar en tiempo real. Igualmente, ENERSUL S.A y la universidad federal de Uberlândia de Brasil [32], utilizaron redes neuronales y árboles de decisión para clasificar el fraude que cometen los consumidores de electricidad con esta compañía.No obstante, los árboles de decisión ayudan a realizar las mejores decisiones sobre la base de la información existente y éste presenta inconvenientes cuando la cantidad de alternativas es grande y cuando las decisiones no son racionales, es decir, el método propuesto carece de capacidad de generalización. También, el instituto de educación Comfenalco desarrolló un proyecto en el cual, mediante encuestas en las que se involucraron diferentes tipos de variables (fisiológicas, sociales, etc), y se clasificó la población estudiantil en edad extraescolar, distinguiendo entre estudiantes con una educación formal y no formal. Esta clasificación se realizó mediante una medida de similitud basada en modas [30], pero los resultados no fueron los esperados, debido a que algunas de las variables utilizadas no aportaron diferencias significativas entre los grupos de clasificación. También, en el año 2013 la universidad autónoma de Puebla presentó

un método para la determinación de los perfiles políticos de una población de votantes, basados en encuestas realizadas y utilizando técnicas de aprendizaje de máquina y teoría de juegos [33]. El propósito fue la elaboración de un modelo que tenga la capacidad de predecir los resultados de las futuras elecciones. El problema principal de este trabajo se debió a que las encuestas no fueron suficientes para determinar los atributos más importantes de los votantes. Lo cual se explica por qué en ocasiones se requieren muchos datos para un buen modelado, y en algunas aplicaciones prácticas, el acceso a grandes cantidades de información no es posible por los elevados costos económicos. Dado el análisis global realizado, puede decirse que hay deficiencias y brechas en el desarrollo de metodologías basadas en aprendizaje de máquina para extraer información relevante e identificar patrones en datos categóricos.

Por otro lado, se han implementado diferentes investigaciones para algoritmos de aprendizaje no supervisado como por ejemplo Ralambondrainy [12], que presenta un enfoque usando el algoritmo K-Means para agrupar datos de tipo categóricos. Este convierte múltiples atributos categóricos en atributos binarios (1 para la presencia y 0 para la ausencia de dicha categoría), luego trata estos atributos binarios como numéricos en el algoritmo K-Means. Sin embargo, este algoritmo convierte todas las características discretas a tributos binarios, aumentado la dimensionalidad de la basa de datos en (2 elevado al número de características), aumentado el costo computacional y el almacenamiento en memoria del algoritmo. Otros algoritmos como el de coeficiente de similitud de Gower [13], medidas de disimilitud, basada en la posición, el alcance y el contenido [19], el algoritmo PAM [8], Clúster jerárquico [20], los algoritmos estadísticos difusos [21] y los métodos de agrupamiento conceptual [22]. Sin embargo, todos estos métodos tienen un problema de eficiencia común cuando son aplicados a una gran cantidad de datos categóricos.

Un acercamiento a la problemática actual fue la de Quian y otros [23] quienes implementaron un algoritmo de agrupamiento el cual se basa en mapear un conjunto de datos categóricos en un espacio euclidiano, su método se basa en la estructura de los datos (SBC) y en sus experimentos obtuvieron un nivel de acierto de clasificación relativamente bueno logrando superar los metodos K-modes [24], Chan [25], Mkm-nof [26] y Mkm-ndm [26], sin embargo este método cuenta con dos problemas, el primero es su elevado costo computacional y el segundo y más relevante, es su pobre nivel de acierto a la hora de agrupar bases de datos categóricas de alta dimensionalidad.

Esta revisión del estado del arte tanto de algoritmos de aprendizaje supervisado y no supervisado y nuevas métricas para datos de tipo cualitativo tiene dos propósitos. Primero, exponer algunas de las falencias que tienen los algoritmos de automatización con los datos categóricos. El segundo, es dejar establecido la necesidad de una nueva medida de disimilitud que permita mapear de un espacio discreto a uno continuo. Esto, para poder mejorar el rendimiento de los algoritmos de agrupamiento, clasificación y regresión.

4.3. Justificación

El mundo de hoy en día evoluciona a pasos agigantados en materia de industria y cada vez se hace más recurrente que las empresas o el sector laborar en general generan grandes cantidades de datos, los cuales, en muchos casos solo quedan almacenados, desaprovechando la oportunidad de generar diferentes tipos de campañas, propuestas etc. Muchos de estos datos son generados a través de encuestas, de internet, etc [2, 29]. Pero, lo común en todos estos datos es que al momento de visualizar la información para poder sacar provecho de ello, se hace imposible, pues todos los datos están traslapados impidiendo incluso un buen acierto de clasificación, regresión o etiquetado [34]. Aun así, una técnica que ha sido comúnmente utilizada para extraer o sintetizar información proveniente de encuestas, es la estadística descriptiva [35]. Pero se conoce que surgen diversos problemas en este enfoque, debido a que los descriptores estadísticos como la media, la desviación estándar, la moda, los cuartiles, etc, generan un sesgo evidente en el análisis [36]. Dado que se introduce correlación entre clases, así como alta dispersión en los datos, el margen de error puede ser elevado, y en ocasiones se hace necesario un tamaño de muestra considerable para lograr resultados concluyentes.

Debido a estos inconvenientes, se conoce que el aprendizaje automático, la inteligencia artificial, en general el mundo computacional está en el auge más grande de la historia y se ha convertido en el día a día del ser humano [37]. Por tal motivo, se debe aprovechar al máximo las facilidades que nos ofrece la tecnología, una forma de hacerlo es adentrarnos dentro del aprendizaje de máquinas más específicamente en bases de datos de tipo categórico [38]. Por ende, la implementación de una nueva medida de disimilitud que permita manejar los datos de tipo cualitativo como datos de tipo cuantitativo sería un gran aporte al mundo del aprendizaje de máquina. Pues, después de mapear de un espacio a otro, sería pertinente utilizar algoritmos estandarizados que existen hoy en día, reduciendo el alto costo computacional y aumentado el rendimiento de los mismos. Implementar la distancia Chi-Square utilizada por algoritmos sofisticados que utiliza el gobierno y demás entidades para el agrupamiento de datos categóricos, será pertinente. Pues, con esto se asegura que la métrica utilizada sería una de las adecuadas para este tipo de datos.

Por ende, sería viable diseñar una metodología que permita mapear datos cualitativos a datos cuantitativos, mejorando la visualización y el acierto de diferentes tipos de algoritmos, pues, sería fundamental para futuros proyectos que se puedan generar en el grupo de Automática de la universidad tecnológica de Pereira, que dependan principalmente de datos cualitativos y que puedan impedir un adecuado manejo en este tipo de datos. Pues, ya se contaría con una herramienta estandarizada y validada con diferentes métricas de rendimiento, que pueda asegurar una alta eficacia y mejorar el rendimiento en todos los algoritmos a utilizar en futuros proyectos.

5. Objetivos

5.1. Objetivo General

Desarrollar una metodología de mapeo de datos de un espacio cualitativo a uno cuantitativo empleando distancias para datos categóricos tal que pueda ser embebida en técnicas de clasificación y agrupamiento.

5.2. Objetivos específicos

1. Desarrollar una nueva medida de disimilitud para datos categóricos utilizando la distancia Chi-square.
2. Desarrollar un enfoque de aprendizaje no supervisado con k-means para validar el rendimiento del algoritmo utilizando la medida de disimilitud ($\mathbf{C} - \mathbf{S}$) .
3. Desarrollar un algoritmo de aprendizaje supervisado mediante tareas de clasificación para validar el rendimiento del algoritmo utilizando la medida de disimilitud ($\mathbf{C} - \mathbf{S}$).

6. Marco teórico

6.1. Embebimiento de la métrica Chi-Square en el algoritmo k-means.

Sea $\mathbf{X} \in \mathbb{Z}^{N\times P}$ un conjunto de datos categóricos con N entradas y P características, nuestro objetivo es encontrar k grupos (clusters) usando el método estándar de K-medias y la distancia Chi-Cuadrado como medida de disimilitud, que es similar a la euclidiana, pero en este caso está ponderada. Esta distancia es una métrica adecuada para el análisis de datos cualitativos, categóricos, nominales y redundantes. Además, compara el recuento de las variables categóricas correspondientes con dos o más características independientes [42]. En consecuencia, consideramos esta distancia como una métrica de disimilitud para mapear datos categóricos en el espacio euclidiano. Construimos la matriz de distancias con la siguiente expresión:

$$d_{ij} = \sqrt{\sum_{n=1}^{P} \frac{1}{\widetilde{w}_n} \left(\widetilde{x}_{in} - \widetilde{x}_{jn}\right)^2}, \tag{1}$$

donde: $\widetilde{x}_{in} = \frac{x_{in}}{\sum_{n=1}^{P} x_{in}}$, $a_n = \sum_{i=1}^{N} x_{in}$ y $\widetilde{w}_n = \frac{a_n}{\sum_{i=1}^{P} a_n}$. En este caso $x_{in} \in \mathbb{Z}$ y $\mathbf{x}_i = \{x_{i1}, ..., x_{iP}\}$ $\in \mathbb{Z}^P$ representa la forma inicial de la muestra categórica, y $\mathbf{d}_i = \{d_{i1}, ..., d_{iN}\} \in \mathbb{R}^N$ es la nueva muestra en el espacio euclidiano, $\widetilde{w}_i \in \mathbb{R}$ puede ser interpretado como i-th peso de las características , de esta manera el conjunto de datos original $\mathbf{X}$ se transforma en un nuevo conjunto de datos $\mathbf{D} \in \mathbb{R}^{N\times N}$. Luego usamos el algoritmo de K-medias que se aplica en $\mathbf{D}$, este método que se usa comúnmente para particionar un conjunto de datos en k grupos (clusters). Esto se hace minimizando la distancia entre muestras del mismo grupo y maximizando la distancia entre objetos que pertenecen a otros grupos [43]. La asignación de grupos se basa en la matriz de distancias, que se calcula con una medida de similitud $\nu\left(\mathbf{d}_n, \boldsymbol{\mu}_k\right)$, y la forma depende de la métrica empleada, siendo $\mathbf{d}_n \in \mathbb{R}^N$ la n-th muestra y $\boldsymbol{\mu}_k \in \mathbb{R}^N$ el k-th centroide [34].

6.2. Incrustación de vecinos estocásticos que siguen una distribución t-student

La incrustación de vecinos estocástica que siguen una distribución t-student **(t-SNE)** Consiste en minimizar la divergencia entre dos distribuciones: una distribución que mide similitudes por pares de objetos de entrada $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) \in \mathbb{R}^{D_1}$ y una distribución

Algorithm 1 El algoritmo básico de K-Means es el siguiente [44]:

1. Inicializar centroides de clúster $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_k \in \mathbb{R}^N$, Aleatoriamente.
2. Repetir hasta que converja:
Para cada i, conjunto:

$$c^{(i)} = \arg\min_j ||\mathbf{d}^{(i)} - \boldsymbol{\mu}_j||^2$$

Para cada j, conjunto:

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{m_k} 1\{c^{(i)}=j\}\mathbf{d}^{(i)}}{\sum_{i=1}^{m_k} 1\{c^{(i)}=j\}}$$

Siendo m_k el número de puntos de datos que pertenecen al k-th grupo c_k

que mide similitudes por pares de los puntos correspondientes de baja dimensión en la incrustación $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N) \in \mathbb{R}^{D_2}$, siendo $D_1 \gg D_2$. Supongamos que nos proporciona un conjunto de datos de objetos de entrada (de alta dimensión) $D = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$ y una función $d(\mathbf{x}_i, \mathbf{x}_j)$ que calcula una distancia entre un par de objetos, por ejemplo, distancia euclidiana $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i – \mathbf{x}_j\|$. Entonces, t-SNE define probabilidades conjuntas $p_{i|j}$ que mide la similitud de pares entre los objetos $\mathbf{x}_i$ y $\mathbf{x}_j$ [45]:

$$p_{j|i} = \frac{exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-d(\mathbf{x}_i, \mathbf{x}_k)^2)/2\sigma_i^2)},$$
$$p_{i|i} = 0$$

$$\sum_{i,j} P_{i,j} = 1$$

Entonces:

$$p_{i,j} = p_{j,i} = \frac{p_{j|i} + p_{i|j}}{2N}$$

En la ecuación anterior, el ancho de banda de los núcleos gaussianos, σ_i, se establece de tal manera que la perplejidad de la distribución condicional p_i s es igual a una perplejidad predefinida μ. Como resultado, el valor óptimo de σ_i varía según el objeto: en las regiones del espacio de datos con una mayor densidad de datos, σ_i tiende a ser más pequeño que en las áreas del espacio de datos. Espacio de datos con menor densidad. El valor óptimo de σ_i para cada objeto de entrada se puede encontrar usando una búsqueda binaria simple [46] o

usando un método robusto de búsqueda.

El objetivo de **t-SNE** es encontrar un D_2 para mapear dimensionalmente $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N) \in \mathbb{R}^{D_2}$ para un reflejo óptimo de las similitudes $p_{i,j}$. Por lo tanto, mide las similitudes $q_{i,j}$ entre dos puntos $\mathbf{y}_i$ y $\mathbf{y}_j$ de manera similar:

$$q_{ji} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_k \sum_{k \neq l} (1 + (||\mathbf{y}_k - \mathbf{y}_l|||)^2)^{-1}}$$
$$q_{ii} = 0$$

Las colas pesadas del núcleo Student-t normalizado permiten que los objetos de entrada diferentes $\mathbf{x}_i$ y $\mathbf{x}_j$ sean modelados por contrapartes de baja dimensión $\mathbf{y}_i$ y $\mathbf{y}_j$ ellos están demasiado alejados Esto es deseable porque crea más espacio para modelar las pequeñas distancias entre pares con precisión (es decir, la estructura de datos local) en la incorporación de baja dimensión. Las ubicaciones del punto de inserción $\mathbf{y}_i$ se determinan minimizando la divergencia de KullbackLeibler entre las distribuciones conjuntas P and Q:

$$C(\varepsilon) = KL(P||Q) = \sum_{i \neq j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$$

Debido a la asimetría de la divergencia Kullback-Leibler, la función objetivo se centra en modelar valores altos de p_{ij} (objetos similares) mediante valores altos de q_{ij} (puntos cercanos en el espacio de incrustación). La función objetivo no es convexa en la incrustación, por lo general, se minimiza usando gradiente descendiente. [47]

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} z (\mathbf{y}_i - \mathbf{y}_j)$$

6.3. Embebimiento de la métrica Chi-Square en el algoritmo t-SNE.

La distancia chi-square es similar a la distancia euclidiana pero ponderada y es la métrica adecuada para el análisis de bases de datos con variables de tipo cualitativas, categóricas

o nominales, datos que se repiten frecuentemente, la distancia chi-square compara los recuentos de respuestas a variables categóricas correspondientes a dos o más características independientes.

$$d_{ij} = \sqrt{\sum_{n=1}^{D} \frac{1}{\tilde{x}_n}(\tilde{x_{in}} - \tilde{x_{jn}})^2}$$

donde

$$\tilde{x}_{in} = \frac{x_{in}}{\sum_{n=1}^{D} x_{in}}$$

$$\tilde{x}_n = \frac{1}{D}\sum_{n=1}^{D} x_{in}$$

Y D es el número de características. La distancia chi-square usa la tabla de contingencia, con la frecuencia de cada atributo (categórico). El ponderado de la distancia **C-S** con las bases de tipo categórico permiten un mejor tratamiento a estos datos. Esto debido a que mejora con creces la separabilidad de la base y permite una mejor visibilidad, permitiendo así realizar agrupamiento y clasificación de una manera mucho más fácil. Sin embargo, uno de los inconvenientes es la alta dimensionalidad que representa mapear los datos a otro espacio de disimilitud, por ende, se decide utilizar el algoritmo **t-SNE** el cual tiene como función reducir la dimensionalidad a 2 o 3 dimensiones. Con el objetivo de conservar la estructura de la base de datos, se decide implementar la matemática de la métrica C-S dentro de la función distancia de **t-SNE**. Esto permite utilizar la **C-S** como distancia y reducir la dimensionalidad y el tiempo computacional con **t-SNE**. [34].

6.4. Técnicas de clasificación estándares

6.4.1. Máquinas de soporte vectorial

Es un algoritmo de aprendizaje supervisado que se puede emplear para clasificación binaria o regresión. Las máquinas de vectores de soporte son muy populares en aplicaciones como el procesamiento del lenguaje natural, el habla, el reconocimiento de imágenes y la visión artificial. Los **SVM** fueron desarrollados por Cortes y Vapnik [39]. Su enfoque puede ser dividido de la siguiente manera:

- Separación de clases: Trata sobre buscar el hiperplano de separación óptima entre las dos clases al maximizar el margen entre los puntos más cercanos de las clases. Los puntos que están ubicados en los límites son los denominados vectores de soporte y la mitad de la margen es el hiperplano de separación óptimo.
- Clases superpuestas:los puntos de datos incorrectos del margen discriminante se ponderan para reducir su influencia (soft margin).
- No linealidad: Cuando no se puede encontrar un separador lineal, los puntos se mapean a otro espacio dimensional en el que los datos se puedan separar linealmente (this projection is realised via kernel techniques).
- Solución del problema: Toda la tarea se puede formular como un problema de optimización cuadrática que puede resolverse mediante técnicas conocidas.

Las máquinas de vectores de soporte pertenecen a una clase de algoritmos de Machine Learning denominados métodos kernel y también se conocen como máquinas kernel [40]:

6.4.2. Clasificador bayesiano

Desde la perspectiva de la probabilidad, según las reglas de Bayes, la probabilidad de que $E = (x_1, x_2, x_3, ..., x_n)$ sea de la clase c es (donde D es el número de atributos o características):

$$p(C|E) = \frac{p(E|C)p(C)}{p(E)},$$

E se clasifica como $C = +$ si y solo si

$$f_b(E) = \frac{p(C = +|E)}{p(C = -|E)} \geq 1,$$

donde $f_b(E)$ se llama clasificador bayesiano. Suponga que todos los atributos son independientes dado el valor de la variable de clase; es decir,

$$P(E|C) = p(x_1, x_2, x_3, ..., x_D|C) = \prod_{i=1}^{D} p(x_i|C),$$

El clasificador resultante es entonces

$$f_b(E) = \frac{p(C = +)}{p(C = -)} \prod_{i=1}^{N} \frac{p(x_i|C = +)}{p(x_i|C = -)}.$$

la funcion $f_b(E)$ esto es llamado el Bayesian Classifier. Donde la diferencia del clasificador discriminante lineal **(LDC)** y cuadrático **(QDC)** es el supuesto de la función de covarianza. Específicamente, si se asume que la covarianza es igual para todas las clases, nos referimos a **LDC**, lo que permite una considerable simplicidad matemática para calcular la distribución de predicción, pero existe una posible pérdida de capacidad de generalización. Si se supone que la covarianza es diferente para todas las clases, nos referimos a **QDC** y podemos separar los datos no lineales con más precisión, pero el cálculo de la distribución de predicción es más complejo. [41].

6.4.3. K-nearest neighbor (K-nn)

El proceso de aprendizaje del método **K-nn** se basa en calcular la distancia del elemento nuevo a cada uno de los existentes, y ordenar dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer. Este grupo será, por tanto, el de mayor frecuencia con menores distancias. El método se describe de la siguiente manera:

- Los datos de entrenamiento $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ con etiquetas $\mathbf{y} = y_1, y_2, ..., y_N$ (siendo N el número de muestras de datos) se almacenan en la memoria.
- Para una nueva muestra $\mathbf{x}_i \in \mathbb{R}^D$, donde D es el número de atributos, se encuentran los k vecinos más cercanos usando una distancia d en todo el conjunto de entrenamiento (k puede ser $1, 3, 5, 7, ...$).
- Se realiza un procedimiento para seleccionar la clase de la nueva muestra. $\mathbf{x}_i$
- las distancias comunes d son:
  - Mahalanobis:
    $$D_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \Sigma^{-1} (\mathbf{x} - \mathbf{y})},$$
    donde Σ^{-1} es la matriz de covarianza entre $\mathbf{x}$ y $\mathbf{y}$.
  - Euclidean:
    $$||\mathbf{x} - \mathbf{y}||_2 = \sqrt{(\mathbf{x} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y})}$$
  - Manhattan:
    $$Manh(\mathbf{x}, \mathbf{y}) = |(\mathbf{x} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y})|$$

7. Algoritmo de agrupamiento usando chi-cuadrado como una distancia

7.1. Introducción

El primer objetivo específico apunta a desarrollar una nueva medida de disimilitud para datos categórico. Este problema representa un desafío constante para investigadores y expertos en el tema como se muestra en el apartado 4. Por lo tanto, se han implementado algoritmos con diferentes medidas de disimilitud para subsanar esta problemática, aumentando en gran medida los niveles de agrupamiento en los algoritmos de aprendizaje de máquina. Sin embargo, el tiempo computacional de estos algoritmos es mayor que la técnica de agrupamiento propuesta en este apartado,que no solo permite mejorar los niveles de acierto en los algoritmos, sino que también realizará procesos con más eficiencia. Como se expresa en el apartado 6 se hizo un adecuamiento de la distribución chi-cuadrada, volviéndola una distancia. Lo que permite entonces que se pueda utilizar dentro del algoritmo kmeans de MATLAB.

7.2. Métodos

La función *kmeans ()* de Matlab en su última versión tiene cinco opciones de medidas de similitud diferentes, que se pueden especificar usando el parámetro 'Distancia'. Sin embargo, ninguna de estas métricas es apropiada para el uso de bases de datos con variables categóricas [48]. Por este motivo, se decide implementar el algoritmo K-Means en Matlab con las métricas estándar, con el agregado de que se incorporó una medida de similitud que no tenía la función, la distancia chi-cuadrado. La principal ventaja que ofrece el algoritmo implementado en este proyecto sobre la función *kmeans ()* de Matlab, es la adición de la distancia chi-cuadrado lo que abre nuevas posibilidades a la hora de agrupar bases de datos con variables cualitativas, como se puede apreciar en el apartado 7.3, la función tiene un correcto funcionamiento comparándose con el método estándar. Por otro lado, la principal desventaja puede ser que, a diferencia de la función *kmeans ()* de Matlab, la función implementada usa el número de iteraciones, así como los centroides, para ir modificando el criterio de convergencia, mientras que la función *kmeans ()* de Matlab usa el algoritmo k -means ++, haciéndolo converger más rápido.

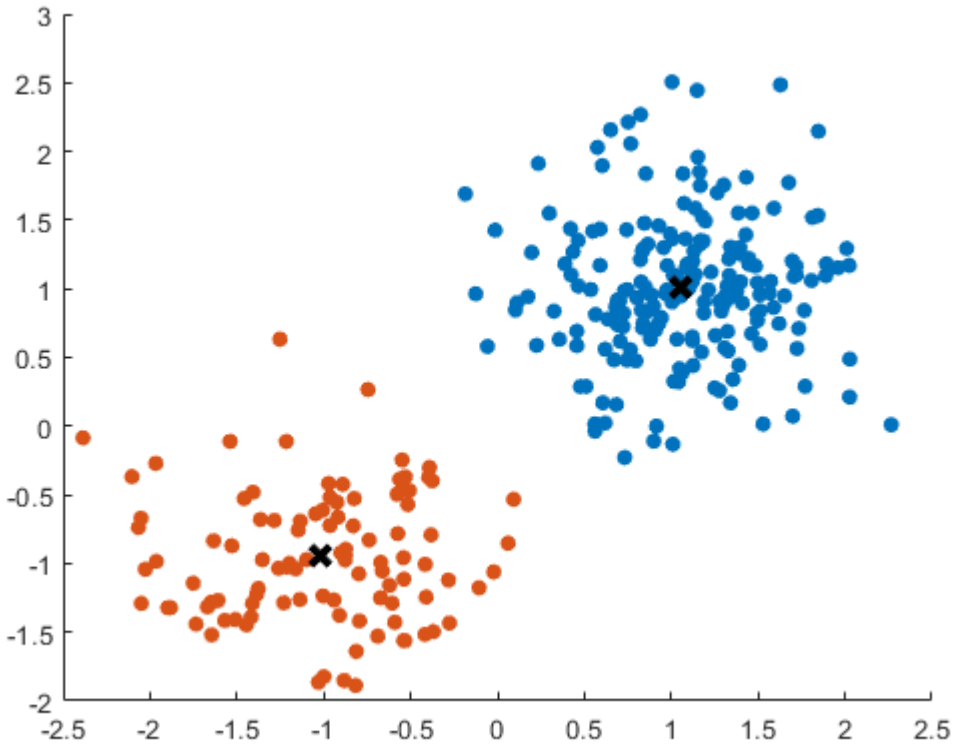


Figura 1: Funcionamiento del algoritmo K-Means implementado, con los parámetros de 10 iteraciones y 2 clusters, aplicado sobre una base de datos sintética.

Cuadro 5: Bases de datos utilizadas en la comparación del algoritmo de agrupación.

Base de datos	Muestras	Características	clases	Tipo de variable
Glass Identification	214	10	6	Cuantitativa
Iris	150	4	3	Cuantitativa
Blood Transfusion Service Center	748	5	2	Cuantitativa
Breast Canceer Wisconsin (Diagnostico)	569	30	2	Cuantitativa
Tic-Tac-Toe Endgame	958	9	2	Categórica
Car Evalution	1728	6	4	Categórica
Congressional Voting Records	435	16	2	Categórica
Balance Scale	625	4	3	Categórica

Para hacer la comparación entre el algoritmo K-Means implementado en este proyecto y la función *kmeans ()*. Se puso a disposición el uso de las bases de datos de UCI Repository [49], de allí se descargaron cuatro bases de datos con variables cuantitativas (donde dos de ellas

son bases de datos biológicas) y cuatro bases de datos con variables de tipo categórico, las cuales son especificado en el siguiente cuadro 5.

7.3. Resultados y discusiones

Poner a prueba el correcto funcionamiento de esta función era prioridad para poder incrustar la nueva distancia chi-cuadrada. Por ende, se tomaron 4 bases de datos cuantitativas y 4 cualitativas, esto, con el objetivo de validar el rendimiento del algoritmo con datos continuos y demostrar la mejora en los datos cualitativos. El experimento tuvo dos etapas, en la primera etapa se probó el algoritmo en las bases de datos cuantitativas, el algoritmo implementado en este proyecto estuvo a la par con la función de Matlab, incluso, mostró un mejor porcentaje de clasificación, aunque se utilizaron las mismas medidas de similitud. En la segunda etapa, los resultados de las pruebas realizadas, muestran como resultado general, un mayor nivel de acierto de clasificación al agrupar con la distancia chi-cuadrado en las bases de datos con variables categóricas, en comparación con las otras distancias.

7.3.1. Etapa 1

Tanto la función implementada en este apartado como la función *kmeans ()* de Matlab fueron probadas en diferentes bases de datos y se calculó el nivel de acierto de los algoritmos al realizar agrupamiento. Para esto, se usaron solo las bases de datos con variables de tipo cuantitativo y un número de 10 iteraciones como criterios de convergencia, además, las medidas de similitud utilizadas fueron sqeuclidean, city block, cosine, and correlation para ambos algoritmos.

La figura 2 muestra la eficiencia de cada algoritmo, donde los porcentajes alcanzados por el algoritmo propuesto teniendo en cuenta la desviación estándar del mismo, es de 10 % mínimo y 55 % máximo. Mientras que el de Matlab, tiene un acierto mínimo de 0 % y un máximo de 42 %. Por otro lado, la figura 3 muestra un mínimo para el algoritmo propuesto de 10 % y un máximo de 27 % y el de Matlab un mínimo de 0 % y un máximo de 23 %. También, la figura 4 muestra un mínimo para el algoritmo de 37 % y un máximo de 64 % y el de Matlab un mínimo de 38 % y un máximo de 61 %. Además, la figura 5 muestra un mínimo para el algoritmo de 20 % y un máximo de 89 % y el de Matlab un mínimo de 20 % y un máximo de 86 %. Esta situación se presenta por la forma en que se elige las $k-medias$ de cada algoritmo.

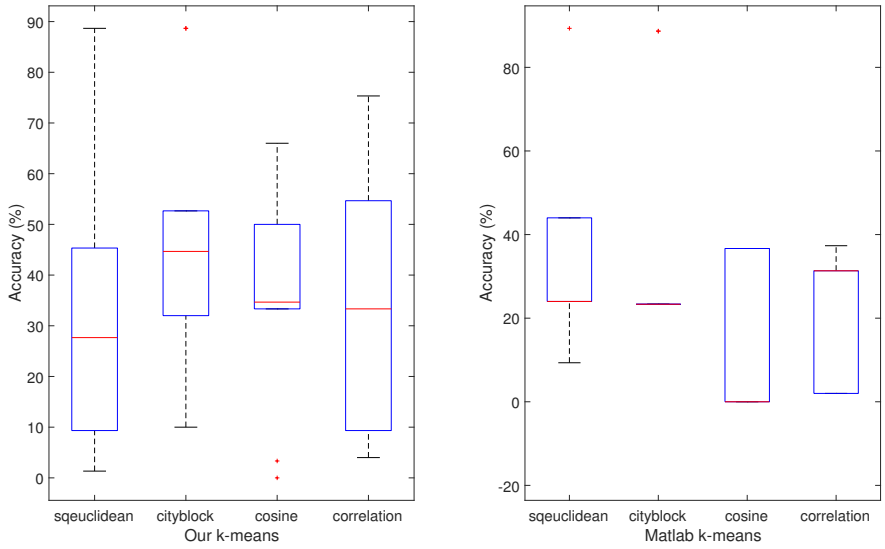


Figura 2: Comparación de los porcentajes de acierto de la base de datos iris donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

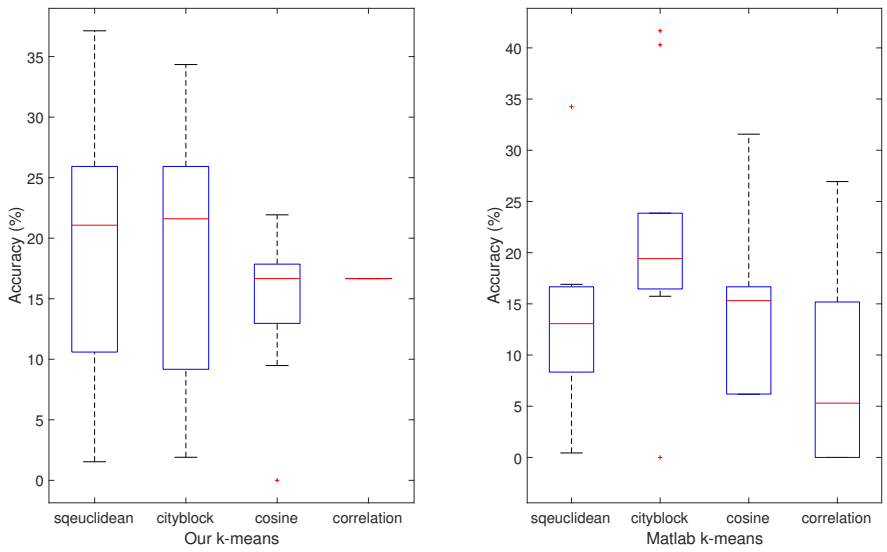


Figura 3: Comparación de los porcentajes de acierto de la base de datos Glass donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

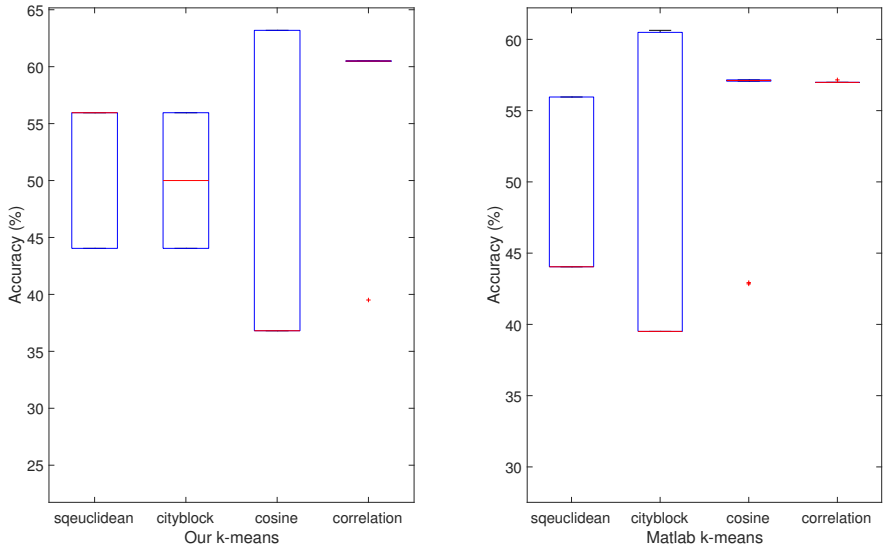


Figura 4: Comparación de los porcentajes de acierto de la base de datos Blood donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

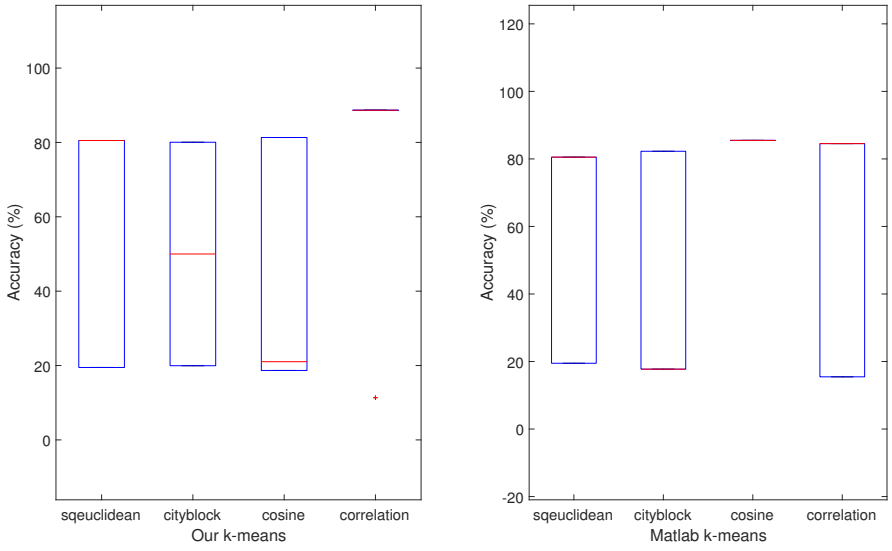


Figura 5: Comparación de los porcentajes de acierto de la base de datos Wisconsin donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

7.3.2. Etapa 2

Finalmente, se comparan ambas funciones usando el nivel de acierto en el agrupamiento, se utilizan las bases de datos con variables de tipo categórico y un número de 10 iteraciones como criterio de convergencia. En esta comparación, la función implementada en este proyecto realizó el agrupamiento basado en la medida de similitud chi-cuadrado, mientras que la función *kmeans ()* de Matlab usó las distancias estándar, excepto la 'correlación' que presentó problemas porque algunos puntos de las bases de datos presentaban desviaciones estándar relativamente pequeñas.

Como se puede observar, la figura 6 muestra el porcentaje de acierto de ambas funciones, en este caso para bases de datos de tipo categórico. El algoritmo propuesto arroja un mayor nivel de acierto utilizando la distancia chi-cuadrado. Basados en la desviación estándar de cada algoritmo, se obtiene un acierto mínimo de 33 % y un acierto máximo de 45 %. Por otro lado, el algoritmo de Matlab usando sus distancias para datos categóricos nos entrega un acierto mínimo de 23 % y un acierto máximo de 43 %. En este mismo orden de ideas, la figura 7 expone para el algoritmo propuesto un acierto mínimo de 49 % y un máximo de 53 % y un mínimo de 44 % y máximo de 56 % para el algoritmo de Matlab. También, en la figura 8 se observa un mínimo de 29 % y un máximo de 44 % para el algoritmo propuesto y un mínimo de 10 % y máximo de 29 % para el de Matlab. Por último, la figura 9 muestra que el algoritmo supera con creces la función de Matlab, mientras el algoritmo propuesto tiene un mínimo de 50 % y un máximo de 88 %, el algoritmo de Matlab muestra un mínimo de 10 % y un máximo de 85 %. Como se observa en todas las figuras, el algoritmo propuesto en este apartado tiene una desviación estándar mas pequeña y la media del acierto siempre se mantiene por encima del promedio estándar de la función de Matlab, lo que permitió concluir que el algoritmo propuesto en este aparatado que esta basado en la distancia chi-cuadrado es viable para variables categóricas.

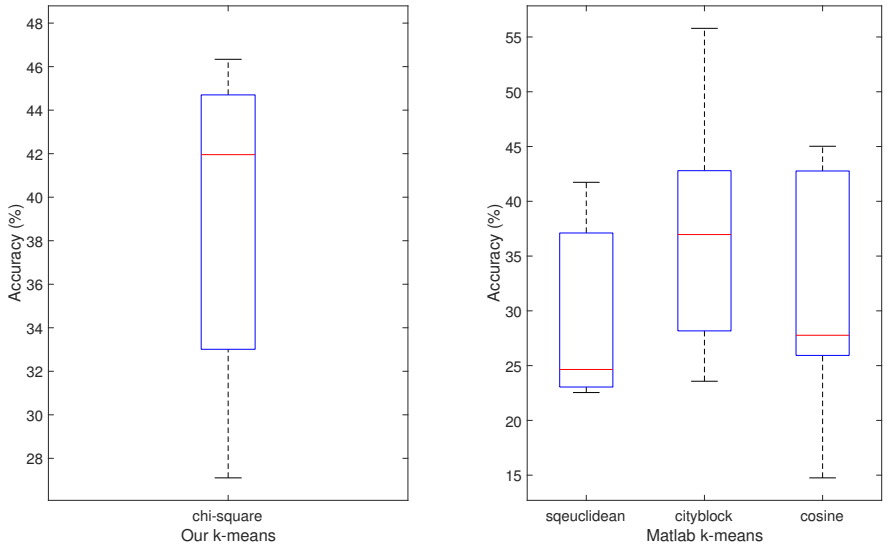


Figura 6: Comparación de los porcentajes de acierto de la base de datos Balance donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

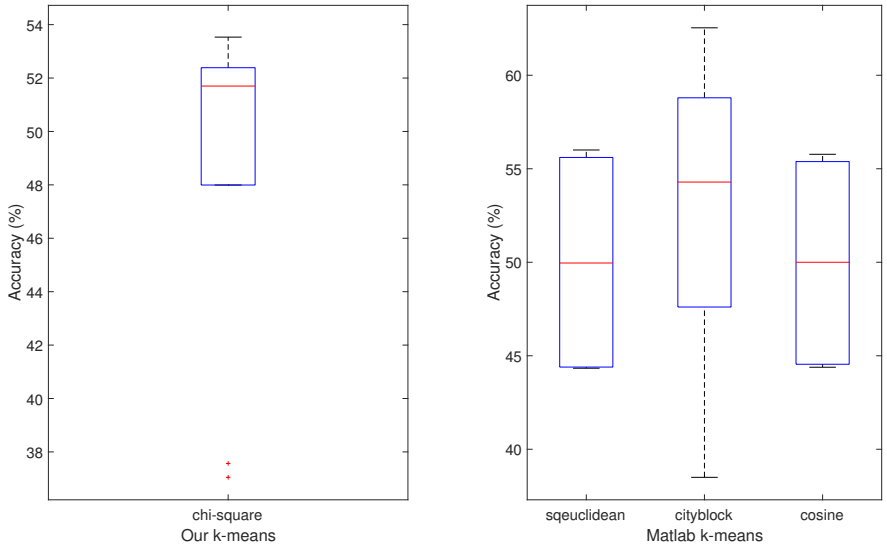


Figura 7: Comparación de los porcentajes de acierto de la base de datos Endgame donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

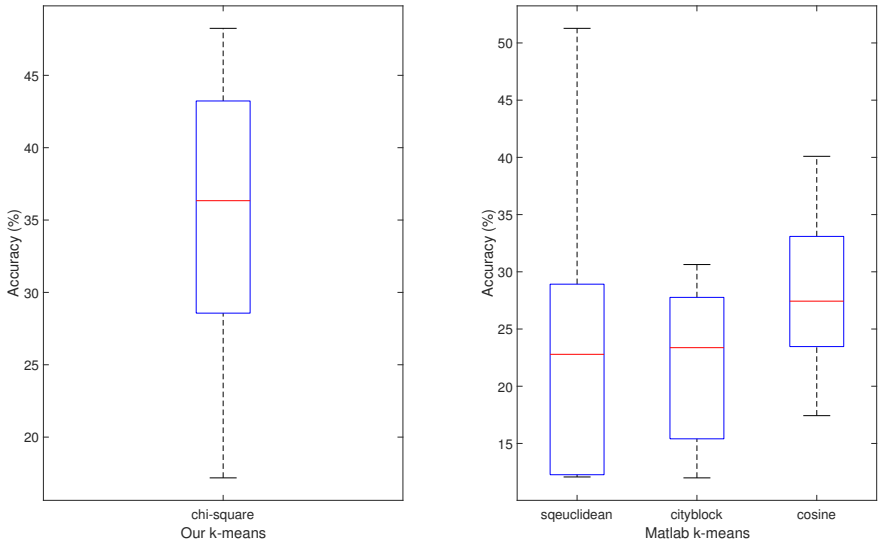


Figura 8: Comparación de los porcentajes de acierto de la base de datos Car donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

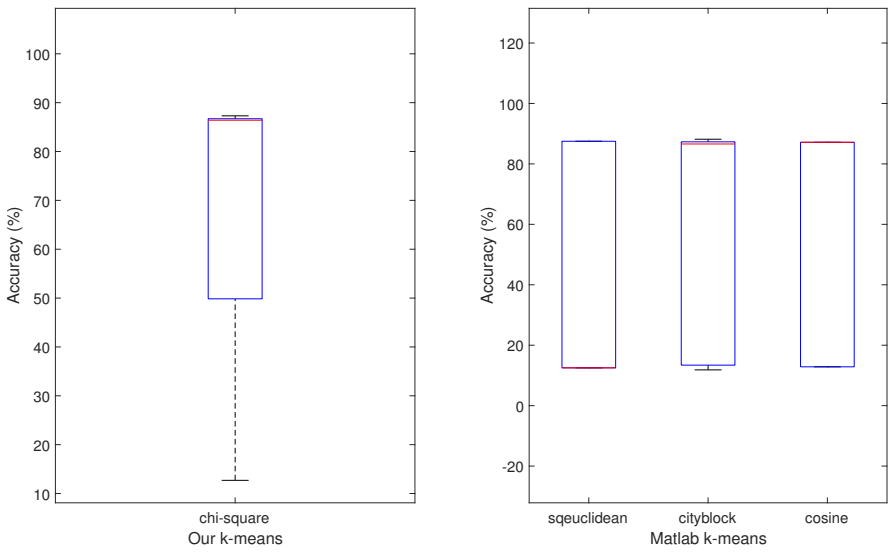


Figura 9: Comparación de los porcentajes de acierto de la base de datos Voting donde: (Our K-means) Es el algoritmo propuesto en el proyecto y (Matlab k-means) es la función de Matlab.

7.4. Resumen

Las diferencias en el porcentaje de éxito entre ambas funciones se pueden atribuir a la forma en que los algoritmos asignan los centroides. La función *kmeans ()* de Matlab selecciona los centroides usando el algoritmo k -means ++. Es decir, seleccione el centroide para la siguiente iteración con una probabilidad proporcional a la distancia de sí mismo al centroide más cercano de la iteración anterior [48]. Mientras que el algoritmo implementado en este proyecto encuentra los centroides en la primera iteración de forma aleatoria y en las otras, lo hace usando la ecuación *kmeans ()* de Matlab. La distancia chi-cuadrado obtiene mejores resultados en los porcentajes de agrupamiento en comparación con las distancias sqeuclidean, cityblock, y cosine. Esto, se puede atribuir a que la distancia chi-cuadrado tiene una ventaja sobre las demás y es que calcula los pesos ponderados de cada característica, asignando una relevancia a cada atributo, lo que hace que esta medida de similitud sea apropiada para bases de datos con variables de tipo categórico.

8. Validación de la medida de disimilitud chi-cuadrado para agrupar conjuntos de datos categóricos

8.1. Introducción

El segundo objetivo específico está dirigido a desarrollar un enfoque de aprendizaje no supervisado con k-means para validar el rendimiento del algoritmo, utilizando la nueva medida de disimilitud ($\mathbf{C} - \mathbf{S}$). La problemática de los datos categóricos, a la cual se le intenta dar solución en el apartado 7, demostró una eficiencia superior a los algoritmos convencionales de Matlab. Sin embargo, validar y demostrar una eficiencia superior ante los métodos de vanguardia para el agrupamiento de datos categóricos se demuestra en este apartado. Como se ha venido mencionando, este algoritmo no solo mejora los niveles de acierto, también, tiene un costo computacional mucho menor como se demuestra en las figuras 15 a 17. Una mejora significativa dentro de este algoritmo es que la métrica chi-cuadrado ahora no se utiliza como una distancia incrustada dentro de kmeans. Ahora se utiliza como una medida de disimilitud, mapeando los datos discretos a un espacio continuo ($\mathbf{C} - \mathbf{S}$) aumentado la dimensionalidad de los datos, lo que permite no solo un aumento en el acierto, también, hace los datos visualmente mas separables como se puede ver en las figuras 12 a 14. Problemática que se ha presentado por décadas en los algoritmos a la hora de visualizar los datos categóricos, pues el traslapamiento de los datos evita que se puedan separar los grupos de forma adecuada.

8.2. Métodos

Sea $\mathbf{X} \in \mathbb{Z}^{N \times P}$ un conjunto de datos categóricos con N entradas y P características, nuestro objetivo es encontrar k grupos (clusters) usando el método estándar de K-medias y la distancia Chi-Cuadrado como medida de disimilitud, que es similar a la euclidiana, pero en este caso está ponderada. Esta distancia es una métrica adecuada para el análisis de datos cualitativos, categóricos, nominales y redundantes. Además, compara el recuento de las variables categóricas correspondientes con dos o más características independientes [42]. En consecuencia, consideramos esta distancia como una métrica de disimilitud para mapear datos categóricos en el espacio euclidiano. Construimos la matriz de distancias con la siguiente expresión:

$$d_{ij} = \sqrt{\sum_{n=1}^{P} \frac{1}{\widetilde{w}_n} \left(\widetilde{x}_{in} - \widetilde{x}_{jn}\right)^2}, \tag{2}$$

donde: $\widetilde{x}_{in} = \frac{x_{in}}{\sum_{n=1}^{P} x_{in}}$, $a_n = \sum_{i=1}^{N} x_{in}$ y $\widetilde{w}_n = \frac{a_n}{\sum_{i=1}^{P} a_n}$. En este caso $x_{in} \in \mathbb{Z}$ y $\mathbf{x}_i = \{x_{i1}, ..., x_{iP}\} \in \mathbb{Z}^P$ representa la forma inicial de la muestra categórica, y $\mathbf{d}_i = \{d_{i1}, ..., d_{iN}\} \in \mathbb{R}^N$ es la nueva muestra en el espacio euclidiano, $\widetilde{w}_i \in \mathbb{R}$ puede ser interpretado como i-th peso de las características , de esta manera el conjunto de datos original $\mathbf{X}$ se transforma en un nuevo conjunto de datos $\mathbf{D} \in \mathbb{R}^{N \times N}$. Luego se usa el algoritmo de K-medias que se aplica en $\mathbf{D}$, este método que se usa comúnmente para particionar un conjunto de datos en k grupos (clusters). Esto se hace minimizando la distancia entre muestras del mismo grupo y maximizando la distancia entre objetos que pertenecen a otros grupos [43]. La asignación de grupos se basa en la matriz de distancias, que se calcula con una medida de similitud $\nu(\mathbf{d}_n, \boldsymbol{\mu}_k)$, y la forma depende de la métrica empleada, siendo $\mathbf{d}_n \in \mathbb{R}^N$ la n-th muestra y $\boldsymbol{\mu}_k \in \mathbb{R}^N$ el k-th centroide [34].

Algorithm 2 El algoritmo básico para K-means se muestra a continuación:

1. Inicializar centroides de clúster $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_k \in \mathbb{R}^N$, Aleatoriamente.

2. Repetir hasta que converja:

Para cada i, conjunto:

$$c^{(i)} = \arg\min_j ||\mathbf{d}^{(i)} - \boldsymbol{\mu}_j||^2$$

Para cada j, conjunto:

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{m_k} 1\{c^{(i)}=j\}\mathbf{d}^{(i)}}{\sum_{i=1}^{m_k} 1\{c^{(i)}=j\}}$$

Siendo m_k el número de puntos de datos que pertenecen al k-th grupo c_k

Primero, se evalúa la métrica adecuada para encontrar los clústeres. Para hacer esto, se evaluaron dos métricas típicas: euclidiana al cuadrado (Sqeuclidean) y coseno, sobre dos conjuntos de datos categóricos: Balance-scale y Tic-Tac-Toe [1]. A continuación, como se señaló antes, se comparó la propuesta K-means ($\mathbf{C} - \mathbf{S}$) con otros métodos de agrupamiento de referencia: SBC [23], K-modes [24], disimilitud ponderada [25], Mkm-nof y algoritmos Mkm-ndm [26]. Se probaron nueve conjuntos de datos públicos descargados de UCI Repository [1] (ver el cuadro 6), evaluando la precisión **(AC)**, el índice de ajuste rand **(ARI)**, la información mutua normalizada **(NMI)** y el tiempo requerido para cada algoritmo. Se validaron todos los métodos en las mismas condiciones: se repitieron los experimentos 100 veces para cada conjunto de datos, y se informaron los valores promedio de **AC**, **ARI** y **NMI** con sus correspondientes desviaciones estándar. Las simulaciones se realizaron utilizando el software Matlab en un servidor Intel (R) Xeon (R), CPU E5-2650 v2 - 2.60GHz, 2 procesadores con 8 núcleos y 280 GB-RAM.

Cuadro 6: Descripción de las bases de datos públicas de UCI Repository. [1]

Base de datos	Muestras	Características	clases	Distribución
Fitting contact lenses (FTL)	24	4	3	$\{4, 5, 15\}$
Ballon (B)	20	4	2	$\{8, 12\}$
Space Shuttle Autolanding (SSA)	15	6	2	$\{6, 9\}$
Soybean-small (SS)	47	35	4	$\{10, 10, 10, 17\}$
Hayes-Roth-Hayes-Roth (HRHR)	132	4	3	$\{51, 51, 30\}$
Lymphography Domain (LD)	142	18	2	$\{81, 61\}$
Vote (V)	435	16	2	$\{168, 267\}$
Breast Cancer (BC)	699	9	2	$\{458, 241\}$
Promoters (P)	106	57	2	$\{53, 53\}$

8.3. Resultados y discusiones

Para definir una distancia para la etapa de agrupamiento, se observa en las figuras 10 y 11 que Sqeuclidean obtiene mejores resultados que la distancia Cosine. Luego, se emplea la Sqeuclidean para todos los métodos de comparación, incluida la propuesta kmeans con chi-cuadrado ($\mathbf{C-S}$). Las figuras 12 a 14 muestran la funcionalidad de la métrica ($\mathbf{C-S}$), que tiene como objetivo aumentar la dimensionalidad de la base de datos, esto, para hacerla visiblemente más separable y obteniendo mejores resultados.Por ende, a la izquierda se encontrará la grafica de la base de datos sin mapear a ningún espacio y a la derecha la base de datos mapeada con la ($\mathbf{C-S}$).

El cuadro 7 muestra los resultados de precisión de los métodos de comparación probados en nueve conjuntos de datos públicos de UCI Repository. El método propuesto muestra los mejores resultados de acierto en la mayoría de las bases de datos. En general, el valor medio de k-means ($\mathbf{C-S}$) es claramente superior al alcanzado por K-modes, W-D, Mkm-nof y Mkm-ndm, y podemos decir que hay una diferencia estadísticamente significativa. En cuanto al método de referencia de la **SBC**, el método propuesto es un poco mejor, pero no hay una disparidad considerable. Los conjuntos de datos categóricos son un tipo de dato complejo porque sus atributos se codifican como valores enteros, lo que genera una gran superposición. Por tanto, los grupos o clases son difíciles de identificar con una precisión aceptable. La distancia Chi-cuadrado permite mapear las características categóricas del espacio euclidiano a uno con una mayor dimensionalidad, sin embargo, con una mejor separabilidad. Por lo tanto, K-means ($\mathbf{C-S}$) reduce el efecto indeseable de superposición y valores atípicos.

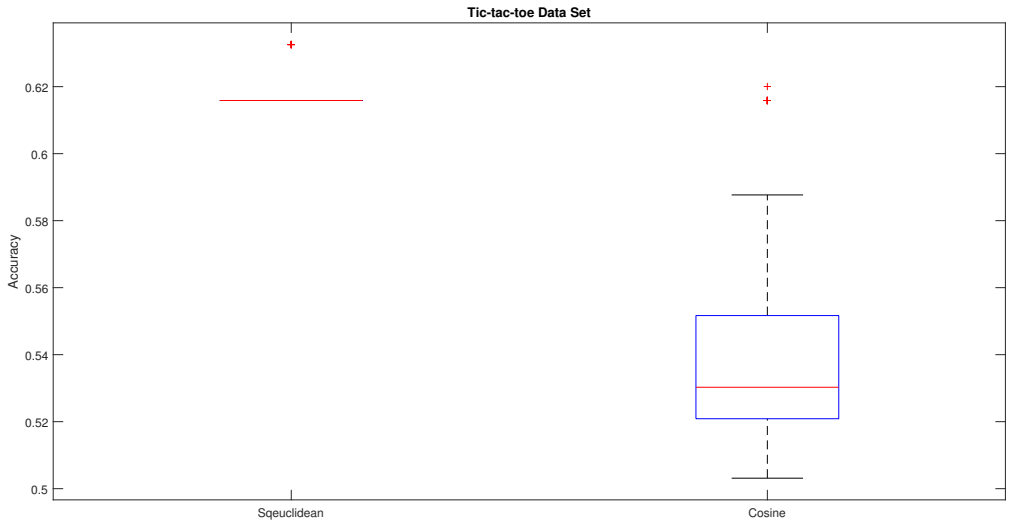


Figura 10: Niveles de acierto para las métricas Sqeuclidean y Cosine en la base de datos Tic-Tac-Toe, para elegir la distancia adecuada del algoritmo k-means.

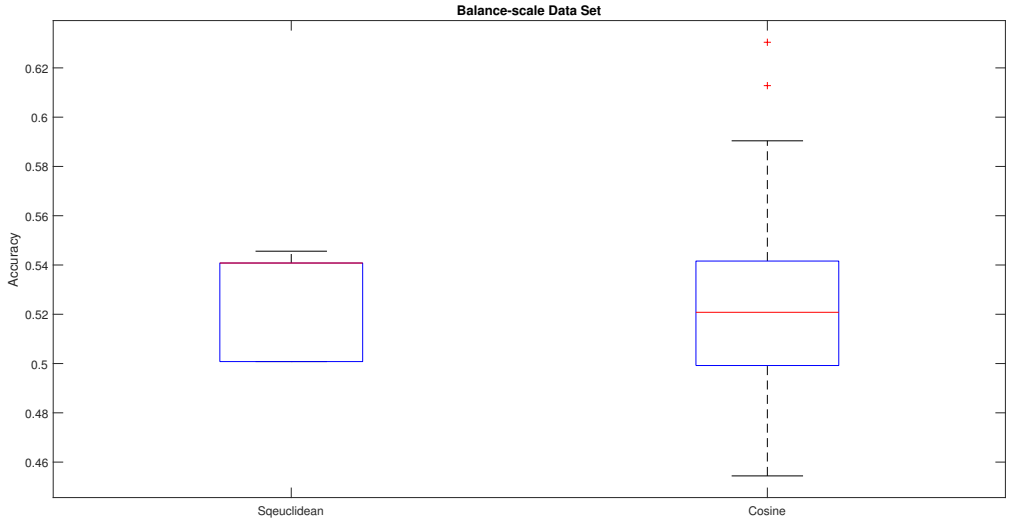


Figura 11: Niveles de acierto para las métricas Sqeuclidean y Cosine en la base de datos Balance-scale, para elegir la distancia adecuada del algoritmo k-means.

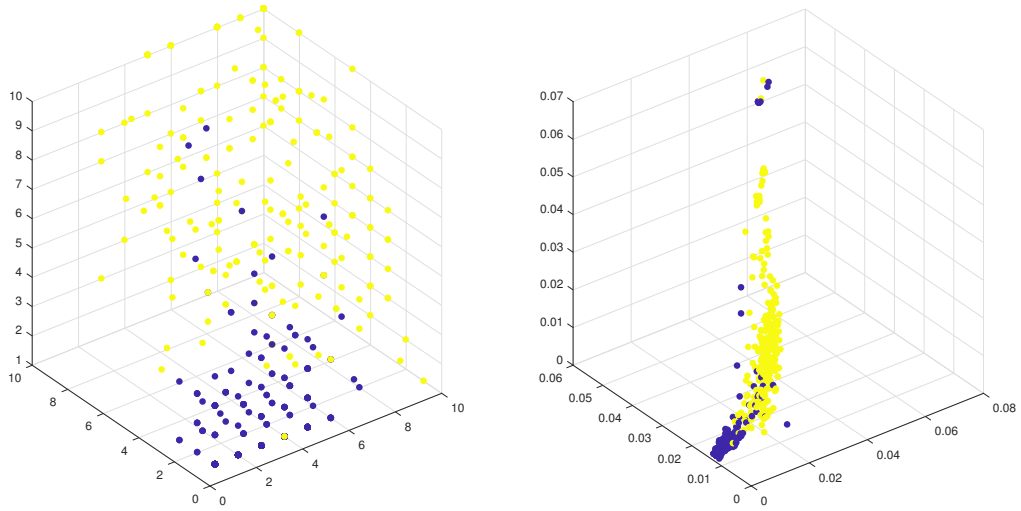


Figura 12: Separabilidad de la base de datos breast cancer: A la izquierda los datos originales y a la derecha los datos mapeados con la distancia $(\mathbf{C} - \mathbf{S})$.

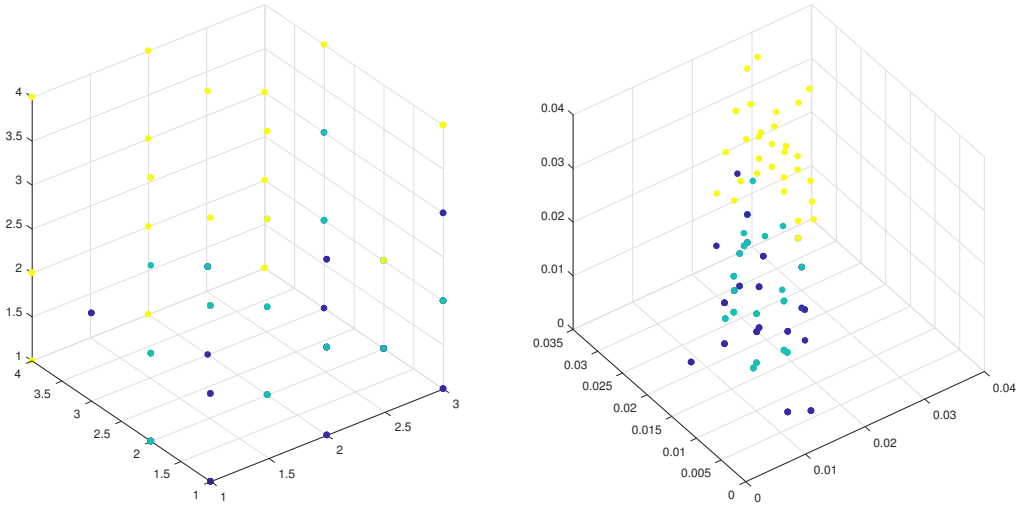


Figura 13: Separabilidad de la base de datos Hayes-Roth: A la izquierda los datos originales y a la derecha los datos mapeados con la distancia $(\mathbf{C} - \mathbf{S})$.

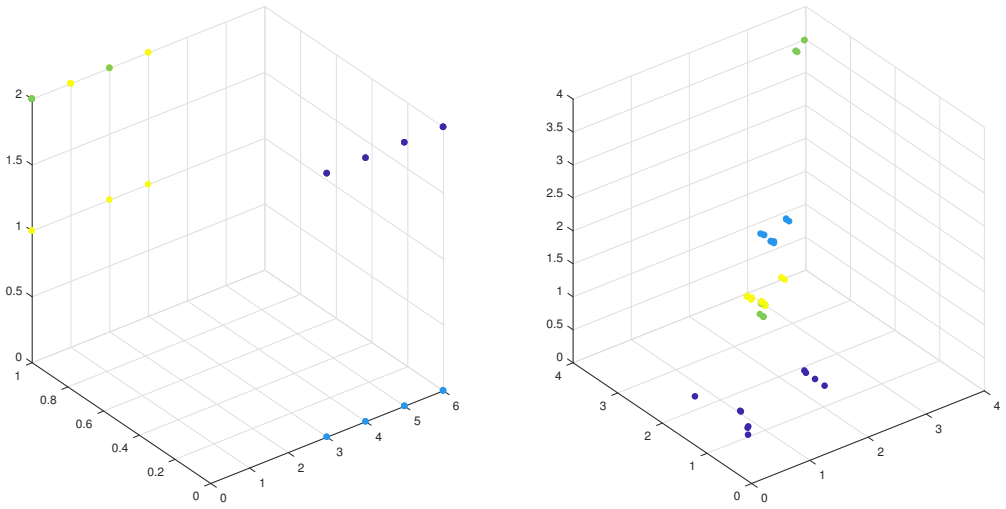


Figura 14: Separabilidad de la base de datos Space Shuttle Domain:A la izquierda los datos originales y a la derecha los datos mapeados con la distancia $(\mathbf{C-S})$.

Cuadro 7: Resultados de precisión de los métodos de comparación probados en nueve conjuntos de datos públicos de UCI Repository. K-means $(\mathbf{C-S})$, se refiere a la propuesta de este apartado, **W-D** corresponde a la disimilitud ponderada. Los conjuntos de datos: FTL, B, SSA, SS, HRHR, LD, V, BC y P se definen en el cuadro 6.

Base de datos	K-means(C-S)	SBC-1	SBC-2	K-modes	W-D	Mkm-nof	Mkm-ndm
FTL	0.6083±0.1046	0.7288±0.0016	**0.7458±0.0005**	0.6417±0.0013	0.6588±0.0016	0.6813±0.0026	0.6587±0.0020
B	**1.0000±0.0000**	0.7940±0.0034	0.8485±0.0001	0.6910±0.0083	0.7045±0.0080	0.7310±0.0070	0.6710±0.0071
SSA	0.6309±0.0959	**0.7140±0.0074**	0.6720±0.0007	0.6240±0.0010	0.6293±0.0011	0.6140±0.0007	0.6367±0.0040
SS	0.8936±0.0000	**0.9660±0.0061**	0.9589±0.0079	0.9185±0.0087	0.8353±0.0096	0.7626±0.0051	0.9483±0.0078
HRHR	**0.7144±0.0788**	0.4566±0.0004	0.4630±0.0009	0.4256±0.0004	0.4782±0.0020	0.4550±0.0019	0.4329±0.0026
LD	**0.8902±0.0125**	0.7618±0.0009	0.7217±0.0008	0.6252±0.0030	0.5808±0.0007	0.5801±0.0000	0.6589±0.0046
V	**0.8852±0.0172**	0.8783±0.0001	0.8759±0.0000	0.8604±0.0001	0.8094±0.0088	0.8715±0.0027	0.8715±0.0027
BC	0.9288±0.0000	0.9293±0.0000	0.9413±0.0000	0.8608±0.0112	0.7717±0.0022	0.7697±0.0000	**0.9464±0.0000**
P	0.7302±0.0520	**0.8878±0.0023**	0.8106±0.0001	0.6335±0.0057	0.7865±0.0028	0.7500±0.0026	0.7043±0.0121
Promedio	**0.8090±0.0401**	0.7907±0.0024	0.7820±0.0012	0.6979±0.0044	0.6949±0.0040	0.6906±0.0025	0.7254± 0.0047

El cuadro 8 muestra los resultados de la métrica de rendimiento **ARI** para cada uno de los método. El metodo propuesto no supera al método **SBC-1**. Sin embargo, tiene un desempeño similar al **SBC-2**, y logra mejores resultados de ARI que los otros métodos de comparación, con diferencias estadísticamente significativas, según una prueba de Kruskal-Wallis. Aunque el método K-means $(\mathbf{C-S})$ y el método **SBC** tienen desempeños similares, el tiempo exigido por el enfoque propuesto es considerablemente menor que el **SBC**.

Cuadro 8: Resultados de comparación de la métrica de rendimiento **ARI** para los diferentes métodos probados en nueve conjuntos de datos públicos de UCI Repository. K-means ($\mathbf{C-S}$), se refiere a la propuesta de este apartado, **W-D** corresponde a la disimilitud ponderada. Los conjuntos de datos: FTL, B, SSA, SS, HRHR, LD, V, BC y P se definen en el cuadro 6.

Base de datos	K-means(C-S)	SBC-1	SBC-2	K-modes	W-D	Mkm-nof	Mkm-ndm
FTL	0.1442±0.0557	0.2897±0.0347	**0.3582±0.0207**	0.0169±0.0088	0.1009±0.0232	0.1232±0.0198	0.0621±0.0118
B	**1.0000±0.0000**	0.3262±0.0215	0.4590±0.0010	0.1356±0.0247	0.1536±0.0253	0.1987±0.0234	0.0981±0.0209
SSA	0.0256±0.0214	**0.1556±0.0248**	0.0566±0.0028	-0.0050±0.0017	0.0064±0.0020	-0.0155±0.0014	0.0262±0.0110
SS	0.7477±0.0000	0.9400±0.0193	**0.9410±0.0159**	0.8247±0.0288	0.6959±0.0214	0.6330±0.0038	0.9111±0.0233
HRHR	**0.2954±0.0566**	0.0133±0.0001	0.0369±0.0005	-0.0018±0.0001	0.0377±0.0010	0.0240±0.0007	0.0071±0.0011
LD	0.0629±0.0141	**0.2721±0.0020**	0.1934±0.0015	0.0627±0.0053	0.0106±0.0011	0.0080±0.0000	0.1109±0.0106
V	0.4470±0.0343	**0.5715±0.0001**	0.5641±0.0000	0.5187±0.0005	0.4123±0.0455	0.5599±0.0127	0.5599±0.0127
BC	0.7369±0.0000	0.7331±0.0001	0.7780±0.0000	0.5395±0.0792	0.2636±0.0126	0.2487±0.0000	**0.7959±0.0000**
P	0.1495±0.0458	**0.6072±0.0082**	0.3802±0.0003	0.0859±0.0065	0.3334±0.0061	0.2545±0.0060	0.2084±0.0310
Promedio	0.3953±0.0253	**0.4343±0.0123**	0.4186±0.0048	0.2419±0.0172	0.2238±0.0153	0.2261±0.0075	0.3089±0.0136

En el cuadro 9 se muestra los porcentajes del índice **NMI**, que es una medida decisiva para calcular la calidad de la agrupación realizada por el método propuesto. Como se puede ver, el método propuesto supera en la mayoría de las bases de datos a los métodos **SBC** demostrando una alta calidad al momento de realizar la agrupación, además, se puede observar en las figuras 15 a 17. Los diagramas de tiempo de ejecución para todos los métodos con una iteración, y se puede observar que el algoritmo propuesto siempre se ejecuta en el menor tiempo, sin importar el conjunto de datos. Para algunos conjuntos de datos, la diferencia en los tiempos de ejecución es muy significativa. En general, estos resultados son relevantes, porque esta propuesta puede aplicarse a cualquier tipo de conjunto de datos, ya sea categórico o cuantitativo. Además, se puede lograr resultados similares al **SBC** robusto, sin embargo, el costo computacional de K-means ($\mathbf{C-S}$) es muy bajo en comparación con métodos similares para agrupar bases de datos categóricas.

Cuadro 9: Resultados de comparación de la métrica de rendimiento **NMI** para los diferentes métodos probados en nueve conjuntos de datos públicos de UCI Repository. K-means ($\mathbf{C-S}$), se refiere a la propuesta de este apartado. Los conjuntos de datos: FTL, B, SSA, SS, HRHR, LD, V, BC y P se definen en el cuadro 6.

Base de datos	K-means(Chi-square)	SBC-1	SBC-2
FTL	**0.2387±0.1073**	0.2102±0.0327	0.2355±0.0924
B	**0.4324±0.3050**	0.3602±0.1648	0.1962±0.0762
SSA	**0.0391±0.0213**	0.0362±0.0047	0.0379±0.0079
SS	0.7949±0.0816	0.7996±0.0668	**0.8330±0.0661**
HRHR	**0.3907±0.0781**	0.0071±0.0149	0.0180±0.0311
LD	**0.0339±0.0229**	0.0283±0.0232	0.0264±0.0309
V	0.4357±0.1381	0.4880±0.0052	0.4898±0.0490
BC	0.4590±0.0049	0.5552±0.0060	**0.7073±0.0000**
P	**0.0733±0.0586**	0.0728±0.0649	0.0761±0.0629
Average	**0.3220**	0.2842	0.2911

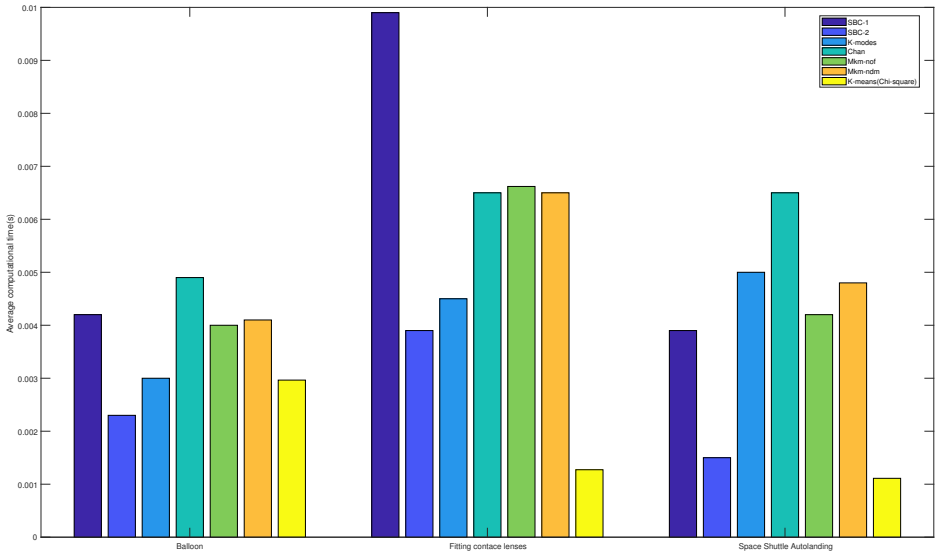


Figura 15: Diagramas de tiempos de ejecución de los algoritmos en las bases de datos Balloon, Fitting contact lenses y Space Shuttle Autolanding.

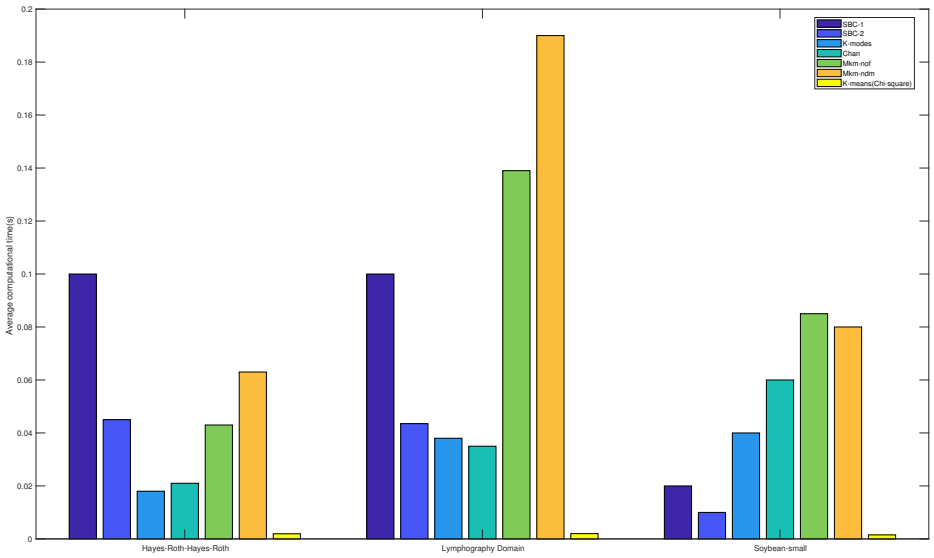


Figura 16: Diagramas de tiempos de ejecución de los algoritmos en las bases de datos Hayes-Roth-Hayes-Roth, Lymphography Domain y Soybean-small.

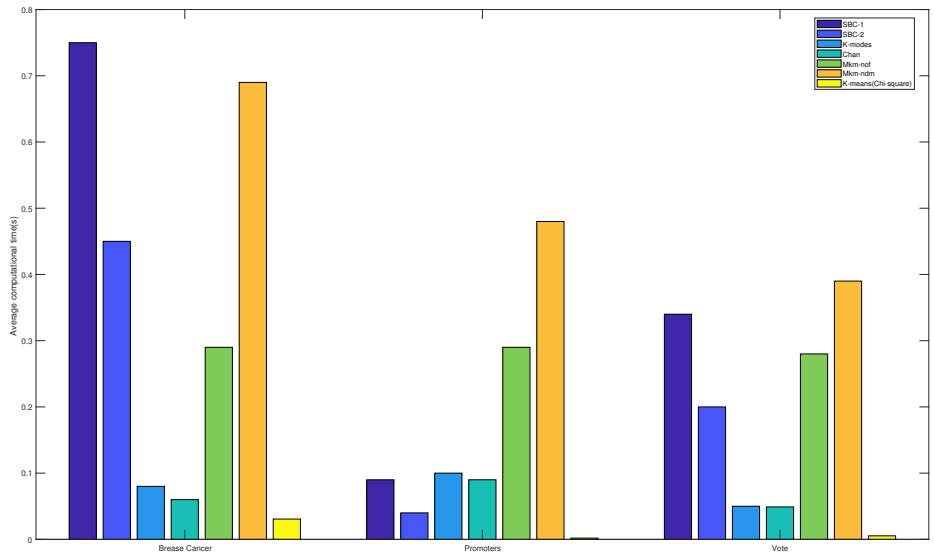


Figura 17: Diagramas de tiempos de ejecución de los algoritmos en las bases de datos Breast Cancer, Promoters y Vote.

8.4. Resumen

En este apartado, se presenta un enfoque alternativo para agrupar conjuntos de datos categóricos. Se adaptó una distancia de disimilitud de chi-cuadrado al método de K-means para mapear las características categóricas de un espacio euclidiano, lo que permite una mejor separabilidad de los grupos o clases. Llamamos a este método K-means ($\mathbf{C-S}$). Esta propuesta mejoró los resultados obtenidos para los métodos de avanzada para la agrupación (K-modes, Mkm-nof, Mkm-ndm) en nueve conjuntos de datos públicos de UCI Repository, y se obtiene un desempeño similar al **SBC** robusto, cuando se evalúa la precisión, el índice rand ajustado y la información mutua normalizada. Además, el K-means ($\mathbf{C-S}$) tiene un costo computacional menor que los métodos de comparación, como se demuestra en los diagramas de tiempo de ejecución de las figuras 15 a 17 Por lo tanto, K-means ($\mathbf{C-S}$) se puede considerar como un método competitivo para agrupar conjuntos de datos categóricos.

9. Clasificación de datos categóricos basados en la medida de disimilitud chi-cuadrado y t-SNE

9.1. Introducción

El tercer objetivo específico está enfocado a desarrollar un algoritmo de aprendizaje supervisado mediante tareas de clasificación para validar el rendimiento del algoritmo utilizando la nueva medida de disimilitud ($\mathbf{C}-\mathbf{S}$). La clasificación es un enfoque interesante para el reconocimiento de datos categóricos. Sin embargo, existen algunos métodos para este propósito en la literatura. Sin embargo, esas técnicas se centran específicamente en los kernels, que tienen problemas de precisión y un alto costo computacional [4]. Por esta razón, se propone un enfoque de identificación de variables categóricas utilizando clasificadores convencionales **(LDC-QDC-KNN-SVM)** definidos en el apartado 1 y una técnica de mapeo para aumentar la separabilidad de clases. Por ende, como se ha venido mencionando, el algoritmo propuesto no solo mejora los niveles de acierto, también, tienen un costo computacional mucho menor como se demuestra en las figuras 15 a 17. Una mejora significativa dentro de este algoritmo es que la métrica chi-cuadrado ahora no se utiliza como una distancia incrustada dentro de kmeans. Ahora se utiliza como una medida de disimilitud, mapeando los datos discretos a un espacio continuo ($\mathbf{C}-\mathbf{S}$) aumentado la dimensionalidad de los datos como se muestra en el apartado 8. Luego, se utiliza el algoritmo **t-SNE** para reducir la dimensionalidad de los datos a 2 o 3 características, lo que entrega un plus a los algoritmos de clasificación, reduciendo el tiempo computacional como se muestra en el cuadro 12. Posteriormente, se evaluó el rendimiento de los diferentes métodos propuestos como se puede ver en la figura 21, donde el método de mapeo propuesto y cualquiera de los 4 clasificadores estándar, mejoran notablemente el acierto de la base de datos sin ningún procedimiento, convirtiendo este método en una forma muy notable en el estado del arte para agrupar y clasificar datos categoricos [55].

9.2. Métodos

9.3. Incrustación de vecinos estocásticos que siguen una distribución t-student

La incrustación de vecinos estocásticos que siguen una distribución t-student **(t-SNE)** Consiste en minimizar la divergencia entre dos distribuciones: una distribución que mide similitudes por pares de objetos de entrada $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) \in \mathbb{R}^{D_1}$ y una distribución que mide similitudes por pares de los puntos correspondientes de baja dimensión en la

incrustación $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N) \in \mathbb{R}^{D_2}$, siendo $D_1 \gg D_2$. Supongamos que nos proporciona un conjunto de datos de objetos de entrada (de alta dimensión) $D = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$ y una función $d(\mathbf{x}_i, \mathbf{x}_j)$ que calcula una distancia entre un par de objetos, por ejemplo, distancia euclidiana $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i – \mathbf{x}_j\|$. Entonces, t-SNE define probabilidades conjuntas $p_{i|j}$ que mide la similitud de pares entre los objetos $\mathbf{x}_i$ y $\mathbf{x}_j$ [45]:

$$p_{j|i} = \frac{exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-d(\mathbf{x}_i, \mathbf{x}_k)^2)/2\sigma_i^2)},$$
$$p_{i|i} = 0$$

$$\sum_{i,j} P_{i,j} = 1$$

Entonces:

$$p_{i,j} = p_{j,i} = \frac{p_{j|i} + p_{i|j}}{2N}$$

En la ecuación anterior, el ancho de banda de los núcleos gaussianos, σ_i, se establece de tal manera que la perplejidad de la distribución condicional p_i s es igual a una perplejidad predefinida μ. Como resultado, el valor óptimo de σ_i varía según el objeto: en las regiones del espacio de datos con una mayor densidad de datos, σ_i tiende a ser más pequeño que en las áreas del espacio de datos. Espacio de datos con menor densidad. El valor óptimo de σ_i para cada objeto de entrada se puede encontrar usando una búsqueda binaria simple [46] o usando un método robusto de búsqueda.

El objetivo de t-SNE es encontrar un D_2 para mapear dimensionalmente $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N)$ $\in \mathbb{R}^{D_2}$ para un reflejo óptimo de las similitudes $p_{i,j}$. Por lo tanto, mide las similitudes $q_{i,j}$ entre dos puntos $\mathbf{y}_i$ y $\mathbf{y}_j$ de manera similar:

$$q_{ji} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_k \sum_{k \neq l} (1 + (||\mathbf{y}_k - \mathbf{y}_l|||)^2)^{-1}}$$
$$q_{ii} = 0$$

Las colas pesadas del núcleo Student-t normalizado permiten que los objetos de entrada diferentes $\mathbf{x}_i$ y $\mathbf{x}_j$ sean modelados por contrapartes de baja dimensión $\mathbf{y}_i$ y $\mathbf{y}_j$ ellos están

demasiado alejados Esto es deseable porque crea más espacio para modelar las pequeñas distancias entre pares con precisión (es decir, la estructura de datos local) en la incorporación de baja dimensión. Las ubicaciones del punto de inserción $\mathbf{y}_i$ se determinan minimizando la divergencia de KullbackLeibler entre las distribuciones conjuntas P and Q:

$$C(\varepsilon) = KL(P||Q) = \sum_{i \neq j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$$

Debido a la asimetría de la divergencia Kullback-Leibler, la función objetivo se centra en modelar valores altos de p_{ij} (objetos similares) mediante valores altos de q_{ij} (puntos cercanos en el espacio de incrustación). La función objetivo no es convexa en la incrustación, por lo general, se minimiza usando gradiente descendiente. [47]

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} z (\mathbf{y}_i - \mathbf{y}_j)$$

9.3.1. Embebimiento de la métrica Chi-Square en el algoritmo t-SNE.

La distancia chi-square es similar a la distancia euclidiana pero ponderada y es la métrica adecuada para el análisis de bases de datos con variables de tipo cualitativas, categóricas o nominales, datos que se repiten frecuentemente, la distancia chi-square compara los recuentos de respuestas a variables categóricas correspondientes a dos o más características independientes.

$$d_{ij} = \sqrt{\sum_{n=1}^{D} \frac{1}{\tilde{x}_n} (\tilde{x_{in}} - \tilde{x_{jn}})^2}$$

donde

$$\tilde{x}_{in} = \frac{x_{in}}{\sum_{n=1}^{D} x_{in}}$$

$$\tilde{x}_n = \frac{1}{D} \sum_{n=1}^{D} x_{in}$$

Y D es el número de características. La distancia chi-square usa la tabla de contingencia, con la frecuencia de cada atributo (categórico). El ponderado de la distancia **C-S** con las bases de tipo categórico permiten un mejor tratamiento a estos datos. Esto debido a que mejora con creces la separabilidad de la base y permite una mejor visibilidad, permitiendo así realizar agrupamiento y clasificación de una manera mucho más fácil. Sin embargo, uno de los inconvenientes es la alta dimensionalidad que representa mapear los datos a otro espacio de disimilitud, por ende, se decide utilizar el algoritmo **t-SNE** el cual tiene como función reducir la dimensionalidad a 2 o 3 dimensiones. Con el objetivo de conservar la estructura de la base de datos, se decide implementar la matemática de la métrica C-S dentro de la función distancia de **t-SNE**. Esto permite utilizar la $(\mathbf{C}-\mathbf{S})$ como distancia y reducir la dimensionalidad y el tiempo computacional con **t-SNE**. [34].

Se probaron siete conjuntos de datos públicos descargados del repositorio de aprendizaje automático de UCI Repository https://archive.ics.uci.edu/ml/index.php. El cuadro 10 describe las bases de datos y sus principales características. Primero, se evaluaron las distancias de **t-SNE** (Cosine, Jaccard, Mahalanobis, Chebychev, Minkowski, City block, Sqeuclidean, Euclidean y la nueva chi-tsne) para demostrar que la métrica $(\mathbf{C}-\mathbf{S})$ combinada con el algoritmo **t-SNE** (chi-tsne) mejora la separabilidad de bases de datos categóricas. Luego, se clasificaron los conjuntos de datos utilizando cuatro enfoques y los clasificadores (LDC, QDC, SVM, K-nn) para encontrar qué método de aprendizaje es el más preciso en este contexto. En aras de la comparación, se probaron cuatro configuraciones diferentes sobre los datos: los clasificadores individuales, los clasificadores + t-SNE, los clasificadores + C-S y los clasificadores + C-S + t-SNE). Consulte el cuadro 4 para obtener una descripción de las configuraciones experimentales. Se calculó la precisión **(AC)** y los tiempos de cálculo para todos los clasificadores en cada configuración, bajo las mismas condiciones. Se realizó un esquema de validación de retención, con diez repeticiones para cada experimento, tomando 70 % de los datos para entrenamiento y 30 % para validación. Las simulaciones se realizaron con el software Matlab en un servidor Intel (R) Xeon (R), CPU E5-2650 v2 - 2.60GHz, dos procesadores con ocho núcleos y 280 GB-RAM.

Cuadro 10: Conjuntos de datos categóricos descargados de UCI Repository.

Base de datos	Muestras	Características	clases	Distribución
Audiology (Standardized) (A)	226	69	2	$\{124, 76\}$
Balloons (B)	16	4	2	$\{12, 8\}$
Breast Cancer (diagnosis) (BC)	699	9	2	$\{458, 241\}$
Chess (King-Rook vs. King-Pawn) (C)	3196	36	2	$\{1669, 1527\}$
Lymphography Domain (LD)	148	18	2	$\{81, 61\}$
Molecular Biology (Promoter Gene Sequences) (MB)	106	57	2	$\{53, 53\}$
Congressional Voting Records (V)	435	16	2	$\{267, 168\}$

9.4. Resultados y discusiones

Las figuras 18 a 20 ilustra el objetivo principal de la $(\mathbf{C}-\mathbf{S})$, que es mapear los datos categóricos a otro espacio para hacerlos más separables. En este caso, se muestran tres de las siete bases de datos (Congressional Voting Records, Balloons and Breast Cancer). se puede ver que el espacio de entrada original (columna de la izquierda) está muy superpuesto y las entidades solo toman valores enteros. Por el contrario, cuando los conjuntos de datos se mapean con el $(\mathbf{C}-\mathbf{S})$, la separabilidad del conjunto de datos aumenta.

El cuadro 11 muestra la precisión y la desviación estándar para **LDC**, **QDC**, **SVM** y **K-nn**, cuando se usa el algoritmo **t-SNE** sobre las bases de datos. El objetivo fue evaluar las distancias (Cosine, Jaccard, Mahalanobis, Chebychev, Minkowski, City block, Seuclidean, Euclidean) comúnmente aplicadas en el método **t-SNE** y demostrar que la $(\mathbf{C}-\mathbf{S})$ es la más adecuada para atributos categóricos. Se puede ver que la métrica $(\mathbf{C}-\mathbf{S})$ incrustada en **t-SNE** supera a las distancias de comparación con diferencias estadísticamente significativas en la mayoría de los casos. Además, el **t-SNE** reduce la dimensionalidad de los datos mapeados sin perder información relevante o la estructura de los datos.

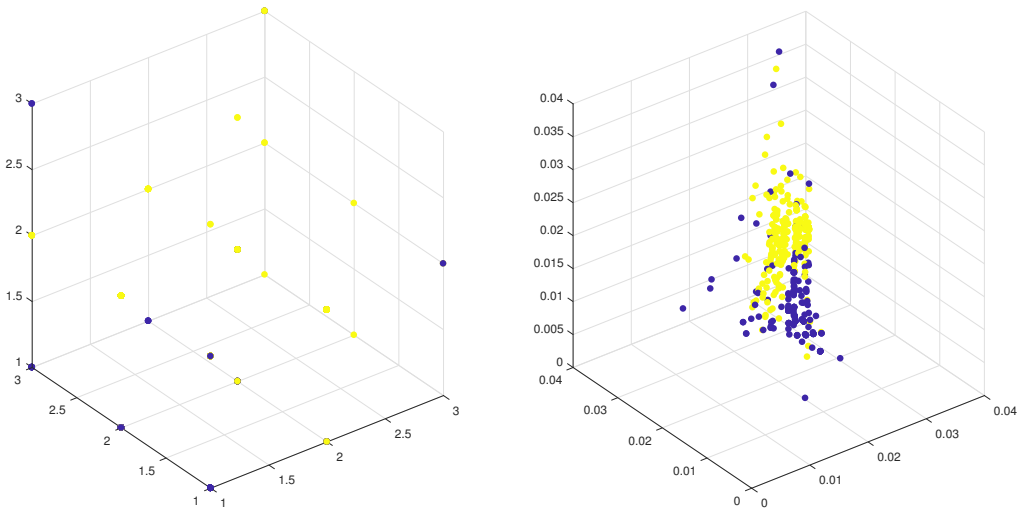


Figura 18: Separabilidad de la base de datos Congressional Voting Records: A la izquierda los datos originales y a la derecha los datos mapeados con la distancia $(\mathbf{C}-\mathbf{S})$.

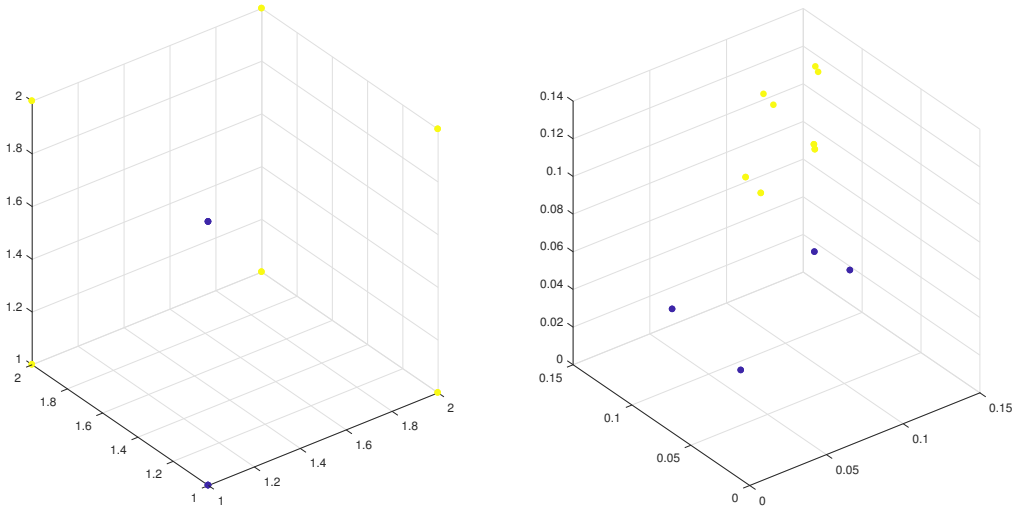


Figura 19: Separabilidad de la base de datos Balloons: A la izquierda los datos originales y a la derecha los datos mapeados con la distancia $(\mathbf{C} - \mathbf{S})$.

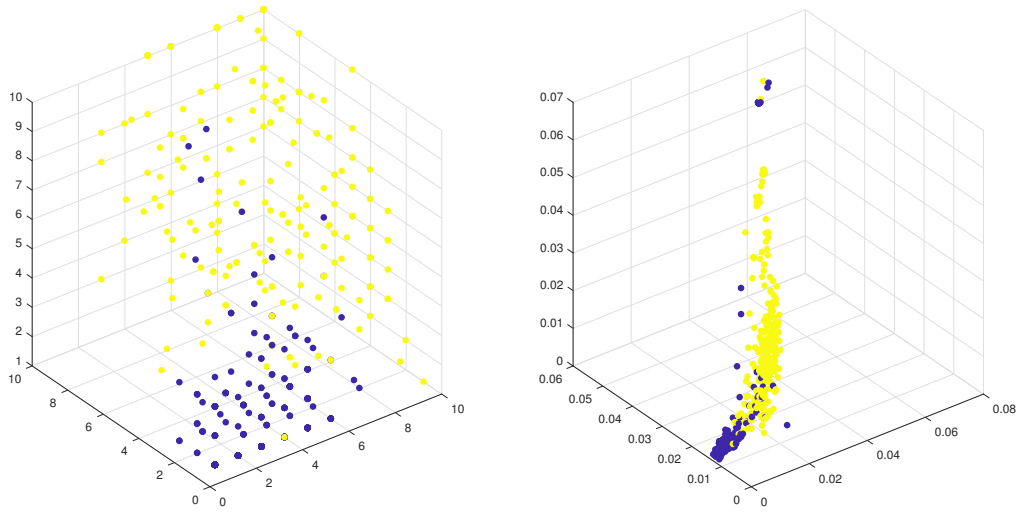


Figura 20: Separabilidad de la base de datos Breast Cancer: A la izquierda los datos originales y a la derecha los datos mapeados con la distancia $(\mathbf{C} - \mathbf{S})$.

Cuadro 11: Resultados de clasificación (acierto) para varias distancias del algoritmo **t-SNE** en siete conjuntos de datos públicos de UCI Respository. LDC y QDC corresponden al clasificador bayesiano lineal y cuadrático, K-nn significa vecino más cercano y SVM es la máquina de vectores de soporte. Los conjuntos de datos: A, B, BC, C, LD, MB, V se definen en el cuadro 10.

Base de datos (A)	cosine	jaccard	mahalanobis	chebychev	minkowski	cityblock	seuclidean	euclidean	**(A) Chi**
LDC	62.5±0.0	70.3±0.1	71.1±0.1	60.0±0.4	69.2±0.0	62.8±0.0	61.2±0.0	66.4±0.0	**73.4±0.1**
QDC	72.8±0.1	83.1±0.0	70.7±0.0	58.5±0.1	73.6±0.3	73.9±0.0	55.6±0.0	69.3±0.0	**84.6±0.0**
K-nn	82.8±0.0	84.8±0.1	77.2±0.0	79.3±0.1	84.4±0.1	85.1±0.0	63.9±0.1	80.8±0.0	**88.9±0.0**
SVM	62.3±0.0	70.8±0.1	71.1±0.0	62.3±0.0	62.6±0.0	60.3±0.0	62.3±0.0	64.3±0.0	**76.7±0.1**
Promedio	70.1	77.2	72.5	65.0	72.5	70.5	60.7	70.2	**80.9**

Base de datos (B)	cosine	jaccard	mahalanobis	chebychev	minkowski	cityblock	seuclidean	euclidean	**(A) Chi**
LDC	74.3±0.2	82.9±0.1	75.7±0.2	78.6±0.2	72.9±0.1	92.9±0.1	54.3±0.1	92.9±0.1	**97.1±0.1**
QDC	71.4±0.2	74.3±0.1	71.4±0.1	75.7±0.2	84.3±0.1	84.3±0.1	75.7±0.2	81.4±0.2	**91.4±0.0**
K-nn	75.7±0.1	85.7±0.1	88.6±0.1	78.6±0.2	85.7±0.1	94.3±0.1	67.1±0.1	95.7±0.0	**100±0.0**
SVM	72.9±0.1	84.3±0.1	85.7±0.2	78.6±0.2	70.0±0.1	94.3±0.1	58.6±0.1	90.0±0.1	**97.1±0.1**
Promedio	73.6	81.8	80.4	77.9	78.2	91.5	63.9	90.0	**96.4**

Base de datos (BC)	cosine	jaccard	mahalanobis	chebychev	minkowski	cityblock	seuclidean	euclidean	**(A) Chi**
LDC	88.3±0.0	94.3±0.0	78.3±0.0	96.3±0.0	95.6±0.0	96.6±0.0	96.5±0.0	96.4±0.0	**96.9±0.0**
QDC	90.6±0.0	93.4±0.0	89.2±0.0	96.8±0.0	96.6±0.0	97.3±0.0	96.5±0.0	96.4±0.0	**97.3±0.0**
K-nn	90.1±0.0	95.3±0.1	91.8±0.0	96.6±0.0	96.7±0.0	97.5±0.0	96.7±0.0	97.1±0.0	**97.4±0.0**
SVM	88.1±0.0	94.4±0.0	79.1±0.0	96.4±0.0	95.5±0.0	96.6±0.0	96.5±0.0	96.5±0.0	**97.2±0.0**
Promedio	89.3	94.4	84.6	96.5	96.1	95.7	96.5	96.6	**97.2**

Base de datos (C)	cosine	jaccard	mahalanobis	chebychev	minkowski	cityblock	seuclidean	euclidean	**(A) Chi**
LDC	60.8±0.0	59.7±0.0	57.8±0.0	50.3±0.0	60.9±0.0	55.3±0.0	62.4±0.0	60.8±0.0	**68.2±0.0**
QDC	65.4±0.0	60.1±0.0	58.9±0.0	53.9±0.0	62.1±0.0	63.1±0.0	64.1±0.0	65.2±0.0	**65.5±0.0**
K-nn	88.5±0.0	70.8±0.0	84.3±0.0	53.0±0.0	89.4±0.0	89.5±0.0	85.9±0.0	89.1±0.0	**89.7±0.0**
SVM	62.6±0.0	60.7±0.0	58.6±0.0	52.2±0.0	61.5±0.0	60.8±0.0	62.5±0.0	61.1±0.0	**68.7±0.0**
Promedio	69.3	62.8	64.9	52.4	68.5	67.2	68.7	69.1	**73.8**

Base de datos (LD)	cosine	jaccard	mahalanobis	chebychev	minkowski	cityblock	seuclidean	euclidean	**(A) Chi**
LDC	76.6±0.0	71.6±0.1	68.9±0.1	64.3±0.1	65.9±0.1	76.1±0.0	76.6±0.0	72.0±0.1	**81.6±0.1**
QDC	77.3±0.1	76.1±0.0	64.1±0.1	67.5±0.1	67.0±0.1	77.5±0.0	78.6±0.0	73.6±0.1	**81.1±0.1**
K-nn	79.1±0.1	76.4±0.1	79.1±0.1	72.5±0.1	74.8±0.1	80.7±0.0	83.4±0.0	78.6±0.1	**84.0±0.1**
SVM	75.0±0.0	71.1±0.1	68.1±0.1	66.1±0.1	68.6±0.1	75.9±0.5	78.9±0.0	70.7±0.0	**81.8±0.1**
Promedio	77.0	73.8	70.0	67.6	69.1	77.6	79.4	73.7	**82.9**

Base de datos (MB)	cosine	jaccard	mahalanobis	chebychev	minkowski	cityblock	seuclidean	euclidean	**(A) Chi**
LDC	47.8±0.1	56.2±0.1	60.0±0.1	43.1±0.1	60.3±0.1	72.5±0.1	62.5±0.1	55.0±0.1	**76.2±0.1**
QDC	57.2±0.1	71.2±0.1	54.1±0.1	57.5±0.1	68.7±0.1	74.7±0.1	65.3±0.1	58.4±0.1	**78.7±0.1**
K-nn	62.5±0.1	70.9±0.1	65.6±0.1	50.9±0.1	66.2±0.0	75.6±0.1	68.1±0.1	70.6±0.1	**80.3±0.1**
SVM	52.2±0.1	55.9±0.1	61.6±0.1	44.4±0.1	56.6±0.1	70.3±0.1	63.7±0.1	54.1±0.1	**76.6±0.1**
Promedio	54.9	63.6	60.3	49.0	63.0	73.3	64.9	59.7	**78.0**

Base de datos (V)	cosine	jaccard	mahalanobis	chebychev	minkowski	cityblock	seuclidean	euclidean	**(A) Chi**
LDC	90.5±0.0	88.9±0.0	90.0±0.0	73.7±0.0	90.2±0.0	91.4±0.0	80.8±0.0	90.1±0.0	**91.5±0.0**
QDC	90.5±0.0	90.9±0.0	90.0±0.0	74.5±0.0	92.1±0.0	91.7±0.0	81.4±0.0	90.6±0.0	**91.4±0.0**
K-nn	92.6±0.0	91.7±0.0	91.4±0.0	76.6±0.0	92.3±0.0	93.3±0.0	82.7±0.0	92.3±0.0	**93.8±0.0**
SVM	91.4±0.0	90.9±0.0	89.8±0.0	75.2±0.0	91.9±0.0	92.3±0.0	81.7±0.0	90.8±0.0	**92.6±0.0**
Promedio	91.2	90.6	90.4	75.0	91.6	92.2	81.6	90.9	**93.1**

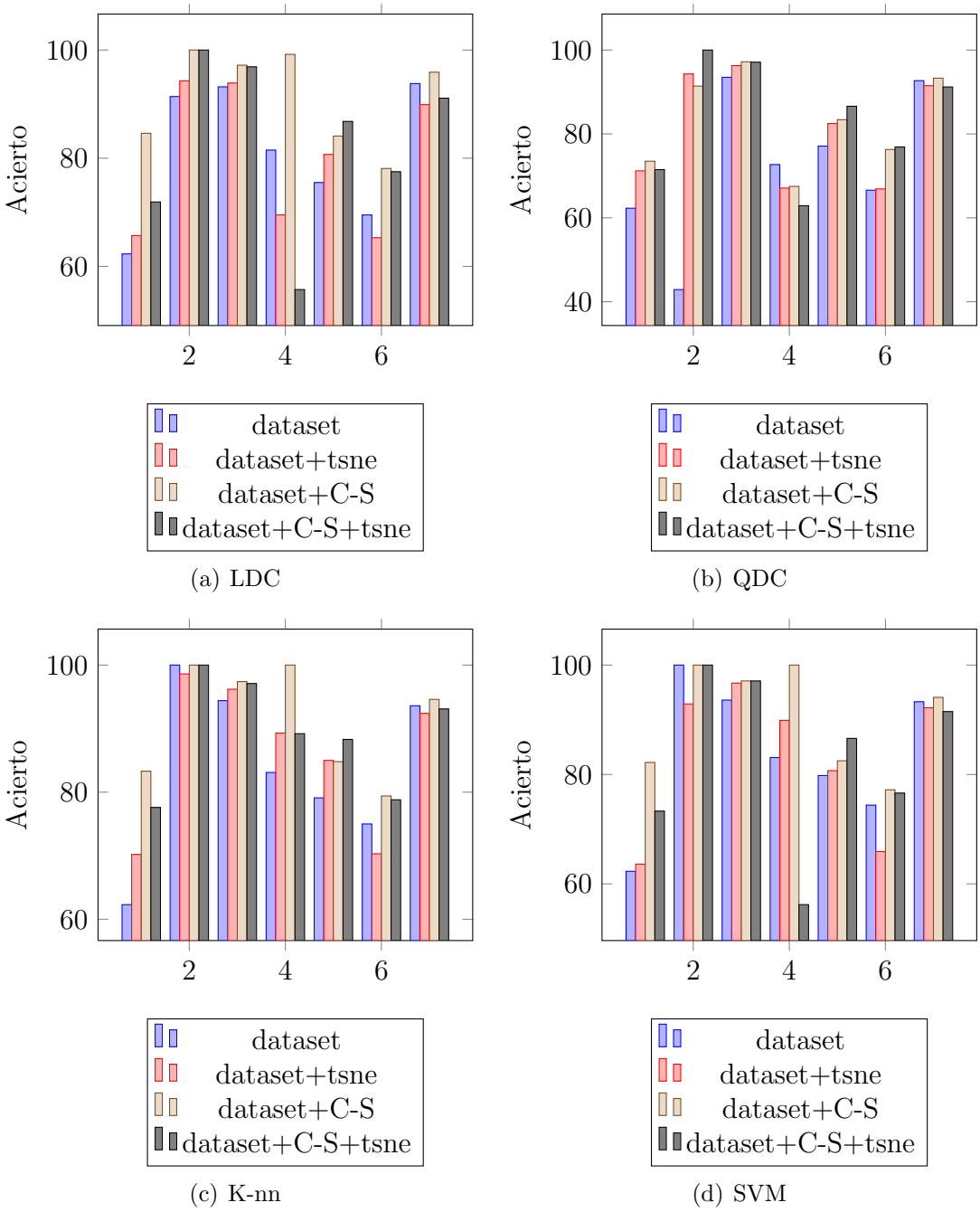
La figura 21 muestra la precisión alcanzada para cada método de aprendizaje en diferentes configuraciones experimentales descritas en el cuadro 4. se pueden identificar cuatro configuraciones diferentes para cada conjunto de datos. El primero, consiste en evaluar los clasificadores estándar en bases de datos categóricas sin ningún procesamiento o mapeo de los datos. Se observa que los resultados de clasificación para datos categóricos no son los mejores para cada conjunto de datos. Esto prueba que los datos categóricos deben procesarse o mapearse antes de las tareas de reconocimiento.

En la segunda configuración, se probaron los clasificadores sobre los conjuntos de datos mapeados con la medida de disimilitud ($\mathbf{C-S}$). Esto permite obtener una mejor separabilidad, pero una mayor dimensionalidad lo que significa mayores tiempos de cálculo. Sin embargo, el mapeo ($\mathbf{C-S}$) genera los mejores resultados de clasificación para todos los conjuntos de datos, como se puede ver en la figura 21. Se considera que este mapeo transforma los datos categóricos en cuantitativos, y los métodos de aprendizaje funcionan mucho mejor en este escenario. Se explica esto de la siguiente manera: la función principal del mapeo ($\mathbf{C-S}$) aumenta la dimensionalidad de los datos para aliviar la superposición de características categóricas. Por esta razón, el mapeo ($\mathbf{C-S}$) realiza una transformación de datos categóricos a cuantitativos.

En la tercera configuración, se realizó una combinación de técnicas de procesamiento. Inicialmente, se mapearon los datos con la medida de disimilitud ($\mathbf{C-S}$). Luego, se aplicó el algoritmo **Chi-tSNE** para reducir el número de atributos a tres dimensiones. Esta reducción de dimensionalidad disminuye los tiempos computacionales mientras preserva la estructura de datos, los resultados de precisión son comparables a los de la segunda configuración, pero los tiempos de cálculo son mucho mejores que los de las otras configuraciones. Esta configuración es la adecuada para sistemas de reconocimiento en línea.

Finalmente, la cuarta configuración aplica **Chi-tSNE** directamente sobre los conjuntos de datos categóricos sin un mapeo ($\mathbf{C-S}$). Aunque los tiempos de cálculo exigidos para entrenar los algoritmos de aprendizaje son menores, la precisión se ve afectada.

En general, se puede ver en La figura 21. que la mejor configuración en términos de precisión fue la segunda, cuando las características categóricas (valores enteros) se mapean con la disimilitud ($\mathbf{C-S}$) a un espacio real (cuantitativo) con mayor dimensionalidad, logrando una mejor separabilidad. Cabe señalar que el mejor clasificador fue el **K-nn** en la mayoría de los experimentos. Es importante mencionar que el método más eficiente en costo computacional fue la tercera configuración como se muestra en el cuadro 12. Esto es notable, porque los porcentajes de precisión son competitivos, además, de lograr los tiempos de cálculo más bajos.



(a) LDC (b) QDC

(c) K-nn (d) SVM

Figura 21: Resultados de precisión **(AC)** de los métodos de comparación probados en siete conjuntos de datos públicos de UCI Repository. Los conjuntos de datos: A, B, BC, C, LD, MB, V se describen en el cuadro 10. Donde a, b, c, d son los clasificadores, donde 1, 2, 3, 4, 5, 6, 7 son las bases de datos. Donde solo la BD es la primera configuración, la BD + t-SNE es la cuarta configuración, la BD + C-S es la segunda configuración, la BD + C-S + t-SNE es la tercera configuración.

Cuadro 12: Resultados del tiempo computacional para los métodos de comparación probados en siete conjuntos de datos públicos de UCI Repository. Los conjuntos de datos: A, B, BC, C, LD, MB, V se describen en el cuadro 10.

Método experimental	**Tiempo computacional (Segundos)**	**Método experimental**	**Tiempo computacional (Segundos)**
(A) + C-S	33.8	**(A) + C-S + t-SNE**	24.6
(B) + C-S	1.4	**(B) + C-S + t-SNE**	0.2
(BC) + C-S	397.0	**(BC) + C-S + t-SNE**	86.0
(C) + C-S	30.9	**(C) + C-S + t-SNE**	21.5
(LD) + C-S	25.9	**(LD) + C-S + t-SNE**	18.3
(MB) + C-S	155.4	**(MB) + C-S + t-SNE**	52.6
(V) + C-S	2530.5	**(V) + C-S + t-SNE**	523.5

Finalmente, como una forma de demostrar la eficiencia de este método, se realizó una comparación de los diferentes métodos de clasificación en las bases de datos categóricas. Después de una exhaustiva revisión del estado del arte [50–53]. se encontraron cinco bases de datos de las siete que se usaron en este trabajo. Teniendo como resultado, un mejor acierto en los algoritmos de clasificación con la metodología propuesta en este proyecto como se puede ver en el cuadro 13.

Cuadro 13: Resultados del acierto de clasificación para la comparación de 4 métodos del estado del arte en 5 de las 7 bases de datos, comparándolos con el método propuesto en este proyecto ($\mathbf{C-S}$).

Database	NBC_c	$C4{,}5$	RIPPER	NPC_c	**C-S**
Chess	87.59±1.23	97.48±1.85	97.22±1.94	88.67±1.72	**100.0±0.00**
Congressional Voting	90.08±3.71	93.28±3.18	92.36±3.23	94.23±3.62	**94.53±1.60**
Breast Cancer	72.50±7.71	71.33±6.33	66.45±6.92	73.81±7.11	**97.35±1.30**
Lymphography Domain	83.60±9.82	73.12±8.63	73.82±8.47	87.76±9.60	**88.30±4.80**
Balloons	100.0±0.00	100.0±0.00	100.0±0.00	100.0±0.00	**100.0±0.00**

9.5. Resumen

En este apartado, se implementó un enfoque de reconocimiento para datos categóricos. Para ello, se desarrollaron dos opciones interesantes: Primero, se asignaron los atributos categóricos a un espacio de mayor dimensionalidad con una disimilitud de Chi-cuadrado ($\mathbf{C-S}$). Este procedimiento permite transformar el dominio de características de conjuntos de datos categóricos de números enteros a valores reales, aliviando el problema de superposición. Se puede observar en la figura 21 que un mapeo de datos categóricos aumenta la precisión en la clasificación. En segundo lugar, Se introdujo una distancia alternativa basada en Chi-cuadrado en el método de incrustación de vecino estocásticos que siguen una distribución

t-student **(tSNE)**; consulte los resultados en el cuadro 11. La combinación de la medida de disimilitud ($\mathbf{C-S}$) y **Chi-tSNE** aplicada a datos categóricos, aumenta simultáneamente la separabilidad de datos y reduce los tiempos computacionales para la clasificación,se probaron los clasificadores estándar: **LDC**, **QDC**, **k-nn** y **SVM** sobre conjuntos de datos categóricos públicos descargados de UCI Repository, como se muestra en el cuadro 12.

El cuadro 13. Muestra, cómo el método propuesto de clasificación usando ($\mathbf{C-S}$) como medida de disimilitud es superior a los otros métodos de clasificación que se usan en cinco de las siete bases de datos propuestas en este proyecto para el apartado de aprendizaje supervisado. Para la comparación en el acierto de la clasificación, se realizó una búsqueda rigurosa en el estado del arte en las siete bases de datos en las que se utilizaron métodos de aprendizaje supervisado [50–53] en la cual se puede observar que el acierto del método propuesto en este proyecto gana con contundencia.

10. Conclusiones

En este trabajo, se presenta un enfoque alternativo para agrupar datos categóricos. adaptando una distancia de disimilitud de chi-cuadrado al método de K-meas para mapear las características categóricas en un espacio euclidiano, lo que permite una mejor separabilidad de los grupos o clases. Se llamó a este método **K-means** $(\mathbf{C}-\mathbf{S})$, propuesta que mejoró los resultados obtenidos por los métodos de avanzada para la agrupación (K-modes, Mkm-nof, Mkm-ndm) en nueve conjuntos de datos públicos de UCI Repository, obteniendo un desempeño similar al **SBC** robusto, al momento de evaluar la precisión, el índice rand ajustado y la información mutua normalizada. Además, el **K-means** $(\mathbf{C}-\mathbf{S})$ tiene un costo computacional menor que los métodos de comparación, como se demuestra en los diagramas de tiempo de ejecución de las figuras 15 a 17. Por lo tanto, **K-means** $(\mathbf{C}-\mathbf{S})$ se puede considerar como un método competitivo para agrupar datos categóricos.

También, se implementó un enfoque de reconocimiento para datos categóricos. Para ello, se desarrolló dos opciones interesantes y adecuadas. Primero, se asignaron los atributos categóricos a un espacio de mayor dimensionalidad con la medidad de disimilitud de Chi-cuadrado $(\mathbf{C}-\mathbf{S})$. Este procedimiento permitió transformar el dominio de las características de los datos categóricos, de números enteros a números reales, aliviando el problema de superposición. Se puede observar en la figura 21, que un mapeo de datos categóricos aumenta el acierto en los algoritmos de clasificación de forma contundente. En segundo lugar, se introdujo una distancia alternativa basada en Chi-cuadrado en el método de incrustación de vecino estocásticos que siguen una distribución t-student **(tSNE)**; cómo se puede observar en el cuadro 11. La combinación de la medida de disimilitud $(\mathbf{C}-\mathbf{S})$ y **Chi-tSNE** aplicada a datos categóricos, aumenta simultáneamente la separabilidad de datos y reduce los tiempos computacionales para la clasificación,utilizando los clasificadores estándar: **LDC**, **QDC**, **k-nn** y **SVM** sobre las bases de datos categóricas descargados de UCI Repository, como se muestra en el cuadro 11.

Para comparar la precisión de los métodos de clasificación, se realizó una búsqueda rigurosa en el estado del arte para las siete bases de datos en las que se utilizaron métodos de aprendizaje supervisado, [50–53] donde se puede observar que el métodos propuesto en este proyecto tienen un mayor acierto de clasificación como se ve en el cuadro 13.

11. Trabajo futuro

Para incrementar el desarrollo de la distancia Chi-square, se proponen los siguientes trabajos a futuro:

- Implementar un análisis de Chi-cuadrado y sus propiedades de invariancia basado en la Matriz de Información de Wasserstein [54].
- Proponer una nueva métrica basada en una formulación de kernel especialmente diseñada para bases de datos cualitativas, por ejemplo, kernels booleanos.
- Desarrollar una formulación de kernel basada en la metrica C-S implementada para datos categóricos.
- Comparar la eficiencia y rendimiento de un kernel para datos categóricos basado en C-S y los kernel convencionales para este tipo de datos.
- Evaluar clasificadores avanzados como los procesos gaussianos profundos o los de aprendizaje profundo.

12. Resultados académicos

Se presentan los artículos publicados y sometidos a congresos y revistas internacionales.

Cuadro 14: Resultados académicos

Objetivo	Categoría	Artículo
Desarrollar una nueva medida de disimilitud para datos categoricos utilizando la distancia Chi-square.	B	Characterization of high school students in the department of Risaralda using the Chi-Square metric. Aceptado para publicación.
Desarrollar un enfoque de aprendizaje no supervisado con k-means para validar el rendimiento del algoritmo utilizando la medida de disimilitud ($\mathbf{C-S}$).	A2	A K-Means Clustering Algorithm: Using the Chi-Square as a Distance [34].
	A2	A Chi-Square dissimilarity measure for clustering categorical datasets. Sometido a **International Journal of Pattern Recognition and Artificial Intelligence**.
	B	Metodología para la predicción de las pruebas saber-11° en los estudiantes de risaralda, utilizando machine-learning. Aceptado para publicación.
Desarrollar un algoritmo de aprendizaje supervisado mediante tareas de clasificación para validar el rendimiento del algoritmo utilizando la medida de disimilitud ($\mathbf{C-S}$).	A2	Classification of Categorical Data Based on the Chi-Square Dissimilarity and t-SNE [55].

13. Agradecimientos

Este proyecto fue financiado por la dirección de investigación y el grupo Organizaciones e innovación de CIAF educación superior, en el proyecto: "Identificación de los atributos más influyentes en los resultados de las pruebas saber 11° en el departamento de la Guajira.". También, este proyecto fue parcialmente financiado por la Maestría en Ingeniería Eléctrica y la Vicerrectoría de Investigaciones, Innovación y Extensión de la Universidad Tecnológica de Pereira.

Referencias

[1] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[2] M. Tascón, *Big Data y el internet de las cosas: qué hay detrás y cómo nos va a cambiar*. Los Libros de la Catarata, 2020.

[3] D. J. Hand, "Principles of data mining," *Drug safety*, vol. 30, no. 7, pp. 621–622, 2007.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[5] M. R. Anderberg, *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Academic press, 2014, vol. 19.

[6] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behavioral science*, vol. 12, no. 2, pp. 153–155, 1967.

[7] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. Oakland, CA, USA, 1967, pp. 281–297.

[8] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[9] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," *Pattern Recognition Letters*, vol. 28, no. 1, pp. 110–118, 2007.

[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[11] J. A. Hartigan, *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[12] H. Ralambondrainy, "A conceptual version of the k-means algorithm," *Pattern Recognition Letters*, vol. 16, no. 11, pp. 1147–1157, 1995.

[13] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

[14] D. Meyer, "Support vector machines. r news, 1 (3): 23–26," 2001.

[15] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*. Springer, 2003, pp. 63–71.

[16] P. K. Janert, *Data analysis with open source tools: A hands-on guide for programmers and data scientists*. .ºReilly Media, Inc.", 2010.

[17] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[18] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446–452, 1999.

[19] K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *pattern recognition*, vol. 24, no. 6, pp. 567–578, 1991.

[20] K. Seshadri and K. V. Iyer, "Design and evaluation of a parallel document clustering algorithm based on hierarchical latent semantic analysis," *Concurrency and Computation: Practice and Experience*, p. e5094.

[21] M. A. Woodbury and J. Clive, "Clinical pure types as a fuzzy partition," *Journal of Cybernetics*, vol. 4, no. 3, pp. 111–121, 1974. [Online]. Available: https://doi.org/10.1080/01969727408621685

[22] R. S. Michalski and R. E. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 4, pp. 396–410, 1983.

[23] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 10, pp. 2047–2059, 2016.

[24] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining." *DMKD*, vol. 3, no. 8, pp. 34–39, 1997.

[25] E. Y. Chan, W. K. Ching, M. K. Ng, and J. Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures," *Pattern recognition*, vol. 37, no. 5, pp. 943–952, 2004.

[26] L. Bai, J. Liang, C. Dang, and F. Cao, "The impact of cluster representatives on the convergence of the k-modes type clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1509–1522, 2013.

[27] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *Journal of artificial intelligence research*, vol. 6, pp. 1–34, 1997.

[28] R. Juste, "La nueva clase dominante: Gestores, inversores y tecnólogos. una historia del poder desde colón y el consejo de indias hasta blackrock y amazon." Arpa, 2020.

[29]

[30] E. Nievas Lio, "Aplicando máquinas de soporte vectorial al análisis de pérdidas no técnicas de energía eléctrica," B.S. thesis, 2016.

[31] D. Gerbec, S. Gašperič, I. Šmon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *IEE Proceedings-Generation, Transmission and Distribution*, vol. 151, no. 3, pp. 395–400, 2004.

[32] E. Gontijo, A. Delaiba, E. Mazina, J. Cabral, J. Pinto *et al.*, "Fraud identification in electricity company customers using decision tree," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, vol. 4. IEEE, 2004, pp. 3730–3734.

[33] I. G. Alonso, M. R. Fernández, J. J. Peralta, A. C. García, and J. M. O. Quintana, "Técnicas de aprendizaje automático al servicio de la eficiencia energética en el hogar digital."

[34] L. A. Serna, K. A. Hernández, and P. N. González, "A k-means clustering algorithm: Using the chi-square as a distance," in *International Conference on Human Centered Computing*. Springer, 2018, pp. 464–470.

[35] H. Oja, "Descriptive statistics for multivariate distributions," *Statistics & Probability Letters*, vol. 1, no. 6, pp. 327–332, 1983.

[36] D. George and P. Mallery, *IBM SPSS statistics 26 step by step: A simple guide and reference*. Routledge, 2019.

[37] J. A. Á. Jareño and V. Coll-Serrano, ""data scientist", today's job." *Métodos de Información*, vol. 9, no. 16, 2018.

[38] S. Nishisato and S. Nishisato, "Analysis of categorical data: Dual scaling and its applications," 1980.

[39] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[40] M. Hu, Y. Chen, and J. T.-Y. Kwok, "Building sparse multiple-kernel svm classifiers," *IEEE Transactions on Neural Networks*, vol. 20, no. 5, pp. 827–839, 2009.

[41] Ş. Büyüköztürk and Ö. Çokluk-Bökeoğlu, "Discriminant function analysis: Concept and application." *Eurasian Journal of Educational Research (EJER)*, no. 33, 2008.

[42] S. Mohanavalli and S. Jaisakthi, "A precise distance metric for mixed data clustering using chi-square statistics," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 10, no. 12, pp. 1441–1444, 2015.

[43] S. Ghosh and S. K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 4, 2013.

[44] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[45] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[46] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *NIPS*, vol. 15. Citeseer, 2002, pp. 833–840.

[47] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[48] 2019. [Online]. Available: https://www.mathworks.com/products/matlab.html

[49] A. Frank, "Uci machine learning repository. irvine, ca: University of california, school of information and computer science," *http://archive. ics. uci. edu/ml*, 2010.

[50] Z. Zheng, Y. Cai, Y. Yang, and Y. Li, "Sparse weighted naive bayes classifier for efficient classification of categorical data," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2018, pp. 691–696.

[51] C. Wang, X. Dong, F. Zhou, L. Cao, and C.-H. Chi, "Coupled attribute similarity learning on categorical data," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 4, pp. 781–797, 2014.

[52] M. Polato, I. Lauriola, and F. Aiolli, "A novel boolean kernels family for categorical data," *Entropy*, vol. 20, no. 6, p. 444, 2018.

[53] K. Baati, T. M. Hamdani, A. M. Alimi, and A. Abraham, "A new classifier for categorical data based on a possibilistic estimation and a novel generalized minimum-based algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 33, no. 3, pp. 1723–1731, 2017.

[54] W. Li and J. Zhao, "Wasserstein information matrix," *arXiv preprint arXiv:1910.11248*, 2019.

[55] L. A. S. Cardona, H. D. Vargas-Cardona, P. Navarro González, D. A. Cardenas Peña, and Á. Á. Orozco Gutiérrez, "Classification of categorical data based on the chi-square dissimilarity and t-sne," *Computation*, vol. 8, no. 4, p. 104, 2020.